The landscape of somatic mutations in primary prostate adenocarcinoma

The landscape of somatic mutations in primary prostate adenocarcinoma

A dissertation presented

by

Sylvan Charles Baca

to

The Division of Medical Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biological and Biomedical Sciences

Harvard University
Cambridge, Massachusetts

April 2013

Advisor: Professor Levi A. Garraway Sylvan Charles Baca

The landscape of somatic mutations in primary prostate adenocarcinoma

Abstract

Prostate cancer is the second leading cause of cancer deaths among men. Targeted analyses of DNA from prostate cancers have identified recurrent somatic alterations that promote tumor growth and survival. Only recently, however, has the comprehensive analysis of cancer genomes become possible due to rapid advances in DNA sequencing technology.

To identify somatic mutations that may drive prostate cancer, we sequenced the protein-coding DNA of 112 prostate tumor/normal tissue pairs enriched for aggressive localized disease. We identified novel recurrent mutations in several genes, including *MED12* and *FOXA1*. The most frequently mutated gene was *SPOP*, which encodes a ubiquitin ligase complex subunit. Mutations altered the substrate-binding cleft of the SPOP protein in 6-15% of tumors across multiple independent cohorts. *SPOP*-mutant prostate cancers lacked ETS gene rearrangements and exhibited a distinct pattern of genomic alterations, including frequent deletion of the chromatin modifying enzyme gene *CHD1*. Transcriptome profiling of prostate epithelial cells suggested that *SPOP* mutations and *CHD1* loss may promote invasive cellular growth and genomic instability, respectively. Thus, *SPOP* mutations appear to define a new molecular subtype of ETS-negative prostate cancer.

In order to characterize the landscape of somatic alterations across the entire genome in prostate cancer, we also sequenced the full complement of DNA from 57 prostate tumors and matched normal tissue. By modeling the genesis of genomic rearrangements, we identified abundant DNA translocations and deletions that arise in a highly interdependent manner. This phenomenon, which we term “chromoplexy”, frequently accounts for the dysregulation of prostate cancer genes and appears to disrupt multiple cancer genes coordinately. Our modeling suggests that chromoplexy may induce considerable genomic derangement over relatively few events in prostate cancer and other neoplasms, supporting a model of punctuated cancer evolution. Together, the studies described in this thesis point toward novel prostate cancer genes and suggest a refined model of prostate tumor evolution.

Table of Contents

Acknowledgements

Many people have guided me through my years of thesis research, as mentors, as friends and as family. I am greatly indebted to those who encouraged me along the way and made my dissertation work possible by giving of their time and expertise.

I have benefited tremendously from the wisdom and guidance of my thesis advisor, Levi Garraway. Dr. Garraway sets an outstanding example as a physician-scientist and as a person that I aspire to emulate. It has been a privilege to learn and grow as a scientist under his guidance. His scientific insight and concern for members of his lab foster a research environment that is wonderfully nurturing and intellectually stimulating.

My colleagues at Dana-Farber Cancer Institute and at the Broad Institute have been invaluable teachers during my foray into the field of cancer genomics. In particular, I thank Judit Jane-Valbuena, Eran Hodis, Michael Lawrence, Yotam Drier, Jean-Philippe Theurillat, Eliezer Van Allen, Franklin Huang and Gad Getz, who have generously shared their knowledge and skills. Similarly, the members of my Defense Advisory Committee, Matthew Meyerson, Wade Harper, Franziska Michor and Massimo Loda, have provided invaluable guidance that has kept my research on track. I am also indebted to my Thesis Defense Committee, including Catherine Wu, Peter Park and Stefano Monti, for generously volunteering their time.

I am deeply grateful to my family for all they have given me. My wife Katie has been a source of unwavering support and encouragement for the last three years. My parents taught me the value of hard work and persistence by their example, not least by toiling to homeschool their five children through elementary school. I owe my love of science to the homeschool days my siblings and I spent cataloguing insects and lizards in our backyard or conducting kitchen table science experiments.

My research endeavors would have been impossible without funding. I am grateful to the NIH Medical Scientist Training Program, which has supported my medical and graduate education (National Institute of General Medical Sciences award number T32GM007753). I also appreciate the countless hours spent keeping me and my fellow MD/PhD students on track by Stephen Blacklow, Marcia Goldberg, Linda Burnley, Robin Lichtenstein, Amy Cohen and the rest of the MD/PhD program staff.

The research described in this dissertation is the result of wide-ranging collaboration with talented scientists within the Harvard community and beyond. Members of the Broad Institute Genome Analysis Platform were instrumental in this work, and performed the DNA sequencing described in this thesis. Several members of Mark Rubin's lab at Weill Cornell Medical College have been wonderful partners in this research. In particular, I thank Chris Barbieri, Kyung Park, Theresa MacDonald, who performed the fluorescence *in situ* hybridization assays and experimental work related to *SPOP* described in Chapter 2. In addition, Francesca Demichelis and Davide Prandi were invaluable collaborators, and in particular spearheaded the analysis of prostate cancer clonal evolution described in Chapter 3.

Lastly, I wish to dedicate this thesis to my friends and family who have battled cancer. Recalling your courage will continue to motivate me in the face of difficulty throughout my career.

CHAPTER 1

Introduction

Adapted from:

Baca, S.C., and Garraway, L.A. (2012). The genomic landscape of prostate cancer. Frontiers in Endocrinology (Lausanne) *3*, 69.

Introduction

Prostate cancer is the second most common malignancy in men and causes over 250,000 deaths each year worldwide (Jemal et al., 2011). At the same time, many men develop indolent prostate tumors that remain asymptomatic and do not require treatment. Thus, prostate cancer stands out in both its prevalence and markedly variable clinical course. Treatments for advanced prostate cancer center on chemotherapy and irradiation as well as hormone ablation therapies that reduce local levels of androgens. While these approaches shrink tumors and alleviate symptoms transiently, metastatic tumors eventually relapse, and often progress rapidly. On the opposite end of the spectrum, overtreatment of benign tumors can cause substantial morbidity from side effects such as impotence and urinary incontinence (Daskivich et al., 2011). Therefore, a deeper understanding of the molecular underpinnings of aggressive prostate cancer may prove valuable, both to distinguish life-threatening disease from benign cases and to elucidate targets for novel therapies.

Like other neoplasms, prostate cancer is driven in part by alterations that accumulate in the DNA of a nascent or growing tumor. Specific mutations of protein-coding genes, gains or losses of gene copies, and chromosomal rearrangements promote the growth of prostate cancers and may differentiate aggressive versus indolent disease. By comparing DNA sequences from a tumor and matched normal tissue from the same individual, one can detect somatic mutations that arose between embryogenesis and removal of the tumor. Many studies have searched in a “targeted” fashion for cancer-associated mutations (reviewed in (Dong, 2006)); however, only in recent years has massively parallel DNA sequencing technology enabled the comprehensive analysis of alterations across the ~3 billion base-pairs in a tumor genome (Meyerson et al., 2010).

The purpose of the research described in this thesis is to leverage recent advances in DNA sequencing technology to identify key alterations that arise in prostate tumor development. From these data, we aim to glean biological insights about the cellular processes and molecular pathways that drive the growth of prostate cancer, including potential nodes for therapeutic intervention. This thesis will describe three lines of research to this end. Following an introduction to the field of prostate cancer genomics in this chapter, Chapter 2 (based on Barbieri et al., 2012) will describe the sequencing all

protein-coding genes ("exomes") from 112 prostate tumors. Through statistical analyses designed to distinguish "driver" mutations from incidental "passenger" mutations, we identified known and novel mediators of prostate tumorigenesis. In Chapter 3 (based on Baca et al., 2013), the analysis of 57 prostate tumor whole genome sequences will be presented. Sequencing entire genomes (rather than the 1-2% of DNA that encodes proteins) allowed us to identify chromosomal rearrangements that often arise in intergenic DNA and to study how they accumulate. We discuss the implications of these findings for tumor evolution and the accumulation of structural alterations in cancer genomes. Lastly, Chapter 4 will discuss the effects on gene transcription of cancer-associated alterations in two genes of interest: the chromatin-modifying enzyme gene *CHD1* and the ubiquitin ligase subunit gene *SPOP*.

To contextualize the research described in this thesis, this chapter (based on Baca and Garraway, 2012) will survey the understanding of prostate cancer genomics prior to this work. We will discuss exemplary somatic mutations in prostate cancer and highlight mutated cellular pathways with biological and possibly therapeutic importance. Examples include mutated genes involved in androgen signaling, cell cycle regulation, signal transduction and development. We will also discuss genetic alterations that may predict the clinical course of disease or response to therapy, as well as the challenges posed to genomic biomarker identification by the molecular heterogeneity of prostate tumors.

The mutational spectrum of prostate cancer

All categories of DNA sequence alterations contribute to prostate tumorigenesis, including point mutations, small insertions or deletions, copy number changes and chromosomal rearrangements (Figure 1.1). An overview of each category of alteration and its contribution to prostate cancer biology, is presented below.

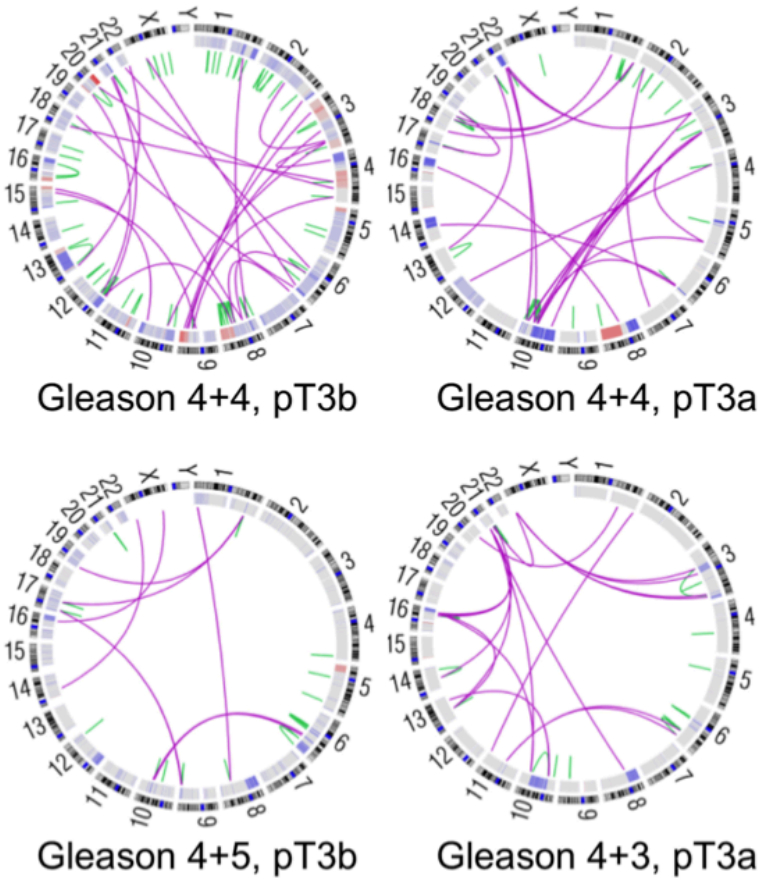


Figure 1.1. Genomic alterations in four high-risk prostate cancers

Circos plot depiction of rearrangements and copy number alterations in four prostate tumor genomes. Intrachromosomal and interchromosomal rearrangements are depicted in green and pink, respectively. Somatic copy number alteration is indicated by red (amplification) and blue (deletion). Gleason scores are listed, indicating the two most prevalent histologic grades in each tumor. Pathological stage is noted as well, where pT3 indicates locally invasive disease.

Somatic copy number alterations

Most prostate cancers exhibit somatic copy number alterations (SCNAs), with genomic deletions outnumbering amplifications in early stages of disease (Visakorpi et al., 1995). Early studies relied on cytogenetics, fluorescence *in situ* hybridiziation and molecular genetic approaches to map candidate cancer genes to regions of SCNA (Brothman et al., 1999). In recent years, comparative genomic hybridization and high-density oligonucleotide arrays have allowed high-resolution analysis of SCNAs across the genome. Statistical analyses of genome-wide copy number data from panels of tumors have pinpointed novel cancer genes in genomic regions that are recurrently deleted or amplified (Beroukhim et al., 2007; Robbins et al., 2011; Taylor et al., 2010).

The extent of SCNA is generally modest in pre-cancerous prostatic intraepithelial neoplasia (PIN), but increases along the spectrum from localized adenocarcinoma to metastatic disease (Zitzelsberger et al., 2001). Particular recurrent SCNAs are enriched in advanced tumors. For example, tumors that fail androgen ablation therapy show frequent amplification of chromosomes 7, 8q and X (Alers et al., 2000; Holcomb et al., 2009; Visakorpi et al., 1995). Animal models of prostate cancer indicate that genes in these regions, such as the androgen receptor gene (X) and the *MYC* proto-oncogene (8q), contribute to cancer progression, as discussed in detail below.

Point mutations and small insertions-deletions

Relative to structural alterations, recurrent point mutations are less common in primary prostate cancers (Kan et al., 2010). Primary tumors generally harbor 1-2 somatic variants per million base pairs – far fewer than known carcinogen-driven tumors such as lung cancer or melanoma, but comparable to breast, renal, ovarian or microsatellite-stable colon cancers (Berger et al., 2011; Greenman et al., 2007; Pleasance et al., 2010a; Pleasance et al., 2010b). While most of these mutations confer no proliferative advantage, a handful of recurrent oncogenic mutations have been defined.

The reported prevalence of mutations in several known cancer genes varies widely and depends on tumor purity, stage, histological grade, and exposure to treatments. For example, *RB1, TP53* and *PTEN* are preferentially mutated in locally advanced or metastatic tumors (Cairns et al., 1997b; Eastham et al., 1995; Tricoli et al., 1996) while the androgen receptor is mutated only in metastatic or treatment-resistant disease (Linja and Visakorpi, 2004; Taylor et al., 2010). Ethnicity may influence mutation prevalence as well. Activating mutations in *KRAS* and *BRAF* occur in ~10% of Asian patients but are rare in Caucasian men, perhaps reflecting different environmental or genetic etiologies of cancers in these populations (Cho et al., 2006; Konishi et al., 1997; Watanabe et al., 1994).

Defects in DNA mismatch repair (MMR) machinery have been reported in prostate cancers and may accelerate progression to castration-independence (Chen et al., 2001; Dahiya et al., 1997). Large-scale sequencing studies have recently identified a subset of tumors with markedly elevated rates of point mutation (Barbieri et al., 2012; Kumar et al., 2011; Taylor et al., 2010). It remains to be determined whether the high levels of mutation in these tumors are caused by MMR deficiency, and whether hyper-mutated cancers display more clinically aggressive behavior.

Structural rearrangements

The discovery of ETS family gene fusions in roughly half of prostate cancers heralded a novel class of alterations in epithelial malignancies as a whole (Tomlins et al., 2005). The most common and prototypical ETS fusion places the oncogenic ERG transcription factor under control of the androgen-regulated *TMPRSS2* gene, leading to high expression in the prostate epithelium. Subsequent research has identified a host of similar oncogenic fusions, where a proto-oncogene is adjoined to a highly active promoter (Kumar-Sinha et al., 2008; Palanisamy et al., 2010; Tomlins et al., 2007). Since mutation or amplification of oncogenes is less common in early-stage prostate cancer, genomic rearrangements may comprise an important means of cancer gene dysregulation in nascent tumors.

Complete sequencing of seven prostate cancer genomes has provided further insight into chromosomal rearrangements in prostate cancer. Primary tumors harbor an average of roughly 100 rearrangements, including translocations, deletions, insertions and inversions (Figure 1.1) (Berger et al.,

2011). Some tumors also display “closed chains” of balanced rearrangements, which appear to arise when multiple DNA breaks occur throughout the genome and the resulting fragments are shuffled and rejoined to one another. These rearrangements may arise when the affected genetic loci are physically proximal to each other, possibly due to co-regulation by transcriptional machinery or nuclear co-localization in open- or closed-chromatin compartments (Berger et al., 2011; Osborne et al., 2004). Consistent with this hypothesis, androgen stimulation can induce physical co-localization of *TMPRSS2* and *ERG* and permit fusion of these genes *de novo* via a topoisomerase 2B-mediate mechanism (Haffner et al., 2010).

The diverse categories of genomic aberrations underscore the need for comprehensive genomic analyses both to understand tumor biology, and perhaps to direct targeted therapies on a genotype-specific basis in the future (Roychowdhury et al., 2011).

Cellular pathways dysregulated by recurrent genomic alterations

Genomic alterations in prostate cancer can increasingly be conceptualized in terms of the molecular processes and pathways on which they impinge (Taylor et al., 2010). Mutations in prostate cancer may affect signal transduction pathways that regulate growth and proliferation, as well as genes involved in the normal development of the prostate. Below, we highlight several themes and pathways that provide a framework for understanding genomic alterations in prostate cancer.

PI3K and MAPK signaling

The phosphoinositide 3-kinase (PI3K) pathway is a central mediator of cellular proliferation and growth that is aberrantly activated in prostate cancer. In response to pro-proliferative signals, PI3K catalyzes the formation of phosphatidylinositol (3,4,5)-triphosphate (PIP_3), which recruits Akt to the plasma membrane. Upon phospo-activation at the plasma membrane, Akt phosphorylates a wide array of substrates that promote proliferation and cell survival.

Prostate tumors achieve activation of PI3K signaling most frequently via inactivation of the tumor-suppressor gene *PTEN* (Figure 1.2). *PTEN* encodes a lipid-protein phosphatase that counteracts signaling by PI3K via dephosphorylation of PIP_3. Loss of heterozygosity at the *PTEN* locus is found in up to 70% of primary prostate cancers and inactivating mutations occur in 5-10% (Barbieri et al., 2012; Cairns et al., 1997b; Gray et al., 1998). Inactivation of PTEN is enriched in advanced tumors and correlates with decreased cancer-specific survival (McMenamin et al., 1999; Sircar et al., 2009). *Pten* disruption in the mouse prostate collaborates with other tumor-promoting events such as inactivation of *Tp53* and overexpression of c-Myc or ERG (Chen et al., 2005; Kim et al., 2011a; King et al., 2009).

Amplification of *PIK3CA*, which encodes the catalytic subunit of PI3K, occurs in 13% to 39% of primary tumors and 50% of castration-resistant tumors (Agell et al., 2011; Edwards et al., 2003; Sun et al., 2009). Activating mutations have been observed in ~5% of primary tumors (Barbieri et al., 2012; Sun et al., 2009). *PIK3CA* activation and *PTEN* loss tend to be mutually exclusive, which suggests functional redundancy—although larger sample sizes are needed to assess this relationship robustly (Sun et al., 2009). Interestingly, *PTEN* loss and *PIK3CA* activation co-occur in other neoplasms such as endometrial cancer, suggesting that these events may engage disparate oncogenic mechanisms in some contexts (Oda et al., 2005). In support of this possibility, oncogenic Akt-independent signaling downstream of mutant *PIK3CA* has been observed in both primary tumors and cancer cell lines (Vasudevan et al., 2009).

The PI3K pathway may be activated by genomic alterations at additional nodes or dysregulated expression of pathway genes (Figure 1.2) (Boormans et al., 2010; Cai et al., 2008; Dong et al., 2006; Taylor et al., 2010). Determining whether these lesions predict sensitivity or resistance to PI3K pathway inhibitors has become an active area of translational research.

The mitogen-activated protein kinase (MAPK) pathway also plays a role in prostate cancer pathogenesis, especially in advanced and castration-resistant tumors. MAPK pathway activation is associated with higher tumor stage and grade and recurrent disease (Gioeli et al., 1999). In the setting of castration resistance, PI3K and MAPK signaling are often coordinately dysregulated (Gao et al., 2006; Kinkade et al., 2008). Evidence for collaboration between these pathways continues to emerge. For instance, PTEN-induced senescence may be overcome by up-regulation of MAPK signaling due to overexpression of HER2 (Ahmad et al., 2011).

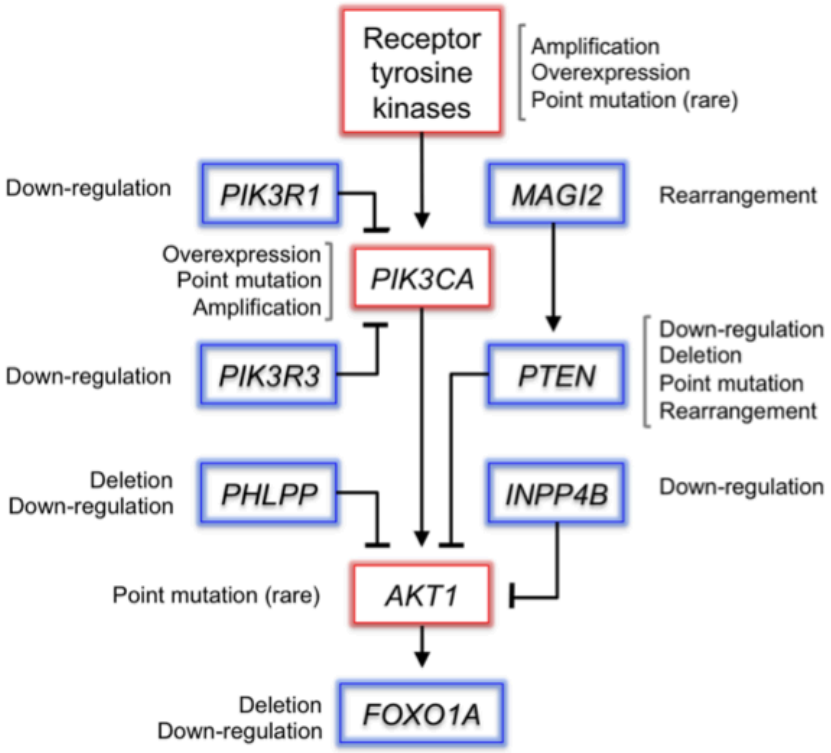


Figure 1.2. Somatic alterations in the PI3K pathway in prostate cancer

Selected members of the PI3K pathway are depicted, alongside the mechanisms by which they are dysregulated in prostate cancer. Putative proto-oncogenes are boxed in red and tumor suppressor genes in blue. The transcript expression of most genes in the pathway is altered (Taylor et al., 2010). The pathway is frequently activated by deletion of *PTEN*. In addition, recurrent deletions inactivate the *FOXO1A* gene, which encodes a transcription factor substrate of Akt that mediates PI3K signaling. *PHLPP1* encodes a phosphatase that dephosphorylates activated Akt, and is frequently co-deleted with *PTEN* in metastatic tumors (Chen et al., 2011). Genomic rearrangements disrupt *MAGI2*, which encodes a scaffolding protein that stabilizes PTEN (Berger et al., 2011; Wu et al., 2000). Although rare, oncogenic point mutations in the receptor tyrosine kinase *EGFR* or *AKT1* may activate the pathway upstream or downstream of PI3K (Boormans et al., 2010; Cai et al., 2008).)

Up-regulation of RAS family members, *RAF1* and *BRAF,* and down-regulation of the counter-regulatory *SPRY1* or *SPRY2* genes are commonly observed in prostate cancer metastases (Kwabi-Addo et al., 2004; McKie et al., 2005; Taylor et al., 2010). In some cases, expression of *RAS*, *RAF1*, and *BRAF* is activated by oncogenic fusion with promoters from highly expressed genes (Palanisamy et al., 2010; Wang et al., 2011). Repression of the RAS-GAP gene *DAB2IP* by EZH2 may activate MAPK signaling and drive progression and metastasis (Min et al., 2010). Defining the relevant mechanisms of pathway activation in greater detail will likely inform strategies for targeting castration-resistant tumors.

Cell cycle regulatory genes

Several cell cycle regulatory genes are disrupted in prostate cancer. Inactivation of cell cycle inhibitors appears to be required to avoid senescence induced by oncogenic signaling and possibly to bypass androgen dependence in metastatic or castration resistant tumors.

The critical cell cycle regulatory genes *TP53* and *RB1* (which encode p53 and Rb) are commonly deleted or mutated in metastatic tumors (Bookstein et al., 1993; Heidenberg et al., 1995; Hyytinen et al., 1999; Tricoli et al., 1996). p53 activates expression of the $p21^{WAF1}$ cyclin-dependent kinase inhibitor. Inactivation of p53 is necessary to bypass cellular senescence mechanisms that are activated upon loss of *PTEN* (Chen et al., 2005). Likewise, Rb regulates the G1 to S cell cycle phase transition, and *RB1* inactivation is particularly common in castration-resistant tumors (Holcomb et al., 2009; Sharma et al., 2010).

Another key cell cycle regulator, *CDKN1B*, encodes the $p27^{Kip1}$ cyclin-dependent kinase inhibitor, and resides within the 12p13 chromosomal region that is frequently deleted. Low $p27^{Kip1}$ expression correlates with poor pathological prognostic markers (Dreher et al., 2004; Vis et al., 2000). Amplification of *SKP2*, which encodes a ubiquitin ligase that targets $p27^{Kip1}$ for proteasomal degradation, may also serve to inactivate $p27^{Kip1}$ (Robbins et al., 2011; Taylor et al., 2010). Disruption of *Cdkn1b* promotes prostate carcinogenesis coordinately with hemizygous deletion of *Pten*, suggesting an interaction between $p27^{Kip1}$ and the PI3K pathway (Di Cristofano et al., 2001). Likewise, the cell cycle regulatory function of $p27^{Kip1}$ induces senescence in PIN lesions driven by Akt1 in mice (Majumder et al., 2008).

Developmental and androgen-regulated genes

Normal developmental and androgen-regulated processes appear to be co-opted during oncogenesis in the prostate. Several genes that participate in the development and differentiation of the prostate epithelium are dysregulated in prostate cancer (Prins and Putz, 2008).

The androgen receptor regulates cellular proliferation and differentiation in response to hormonal signals in the prostate epithelium. While androgen receptor mutations are absent in primary tumors, the *AR* gene is frequently mutated or amplified in metastatic and castration-resistant disease (Koivisto et al., 1997; Linja and Visakorpi, 2004; Visakorpi et al., 1995). Castration resistant tumors remain dependent on the androgen receptor, and overexpression of AR confers castration-resistant growth (Chen et al., 2005). AR point mutations allow promiscuous activation by steroid hormones such as estrogens, progestins, glucocorticoids and androgen antagonists in 10-30% of refractory cases (Gaddipati et al., 1994; Linja and Visakorpi, 2004). Dysregulation of androgen signaling may contribute to localized disease as well: mutation and altered expression of several AR-interacting genes including *NCOR2*, *NRIP1*, *TNK2* and *EP300* are observed in both primary tumors and metastases (Taylor et al., 2010).

NKX3-1 encodes a prostate-specific transcription factor that is required for normal development of the prostate and is deleted or down-regulated in up to 90% of prostate cancers (Asatiani et al., 2005; Emmert-Buck et al., 1995; Vocke et al., 1996). Inactivation of this gene via hemizygous deletion of chromosome 8p appears to occur early and can be observed in PIN lesions (Asatiani et al., 2005; Emmert-Buck et al., 1995). *Nkx3-1*-deficient mice exhibit defective branching morphogenesis of the prostate gland and develop PIN-like lesions with age (Bhatia-Gaur et al., 1999). In addition, NKX3-1 appears to protect the differentiated prostate epithelium from oxidative DNA damage (Bowen and Gelmann, 2010; Ouyang et al., 2005). Therefore, loss of *NKX3-1* may both disrupt terminal differentiation and foster the mutational inactivation of collaborating cancer genes such as *PTEN* (Kim et al., 2002).

The Wnt pathway regulates embryological development, and its contribution to prostate cancer is becoming increasingly recognized (Yardy and Brewster, 2005). Key pathway genes including *APC*, *AXIN1* and the β-catenin gene *CTNNB1* may be mutated at low frequency (Chesire et al., 2000; Voeller et

al., 1998; Yardy et al., 2009). *APC* undergoes LOH in roughly 20% of primary cancers and promoter CpG methylation in up to 90% (Brewster et al., 1994; Phillips et al., 1994; Yegnasubramanian et al., 2004). β-catenin may promote proliferation through co-activation of AR-mediated transcription (Cronauer et al., 2005; Truica et al., 2000). Additional mutations in Wnt pathway genes were recently documented in the progression to castration-resistant disease (Kumar et al., 2011). More pairs of pre- and post-relapse samples should be analyzed to clarify the importance of this pathway in refractory disease.

Genomic heterogeneity of prostate cancer

Prostate cancer is a clinically and genetically heterogeneous disease. Independent cancerous foci with distinct morphological features often coexist in a single prostate. The course of disease also varies widely: some cancers remain indolent for decades while others rapidly progress to lethality. Distinct molecular features appear to underlie the clinical and histological differences. Identifying genomic determinants of aggressive disease might improve experimental modeling and stratification of patients with intermediate-risk prostate cancer.

Prostate cancer may arise in multiple foci from independent precursor cells that are driven to neoplastic transformation by carcinogenic exposures or genetic predisposition (Andreoiu and Cheng, 2010). The presence of genomic lesions can vary between foci in a single tumor, including *TMPRSS2-ERG* fusion, *MYC* amplification and *TP53* mutation (Jenkins et al., 1997; Mehra et al., 2007; Mirchandani et al., 1995). Multiple distinct clones can be identified in a single biopsy (Ruiz et al., 2011), but most metastatic prostate cancers appear to originate from a single clone within a primary tumor (Holcomb et al., 2009; Liu et al., 2009b; Qian et al., 1995). Among other lesions, subclonal *TP53* mutations may define cells in the primary tumor with metastatic potential (Mirchandani et al., 1995; Navone et al., 1999). Intra-tumoral heterogeneity complicates efforts to define prognostic mutations or expression signatures from primary tumors, because the primary tumor subclone that gives rise to metastatic disease must be adequately sampled (Sboner et al., 2010).

Despite the challenges posed by tumor heterogeneity, expression signatures have been proposed that delineate histologically aggressive disease or predict outcome independently of clinical variables (Febbo, 2009; Glinsky et al., 2004; Singh et al., 2002; True et al., 2006). However, the overlap between signatures from independent studies is moderate. Some genomic alterations appear to have prognostic value as well. The *TMPRSS2-ERG* fusion, *MYC* amplification, and *PTEN* or *p53* deletion predict cancer-specific death in at least some patient cohorts (Demichelis et al., 2007; Sato et al., 1999; Sircar et al., 2009). In some cases, a mutational signature may underlie expression-based sub-classifications (Lapointe et al., 2004; Lapointe et al., 2007).

Conclusion

In summary, the analysis of prostate cancer genomes using targeted sequencing approaches and gene copy number profiling has pointed toward molecular pathways and processes that are dysregulated in prostate cancers. Disrupted genes include mediators of androgen signaling and prostatic development as well as regulators of cell division. Importantly, genomic alterations target both generic cancer-associated pathways and processes that impinge more specifically on prostate tissue. Therefore, the unbiased characterization of prostate cancer genomes through whole-exome and whole-genome sequencing may hold promise for the discovery of novel cancer genes that have not previously been identified in other cancer types. Along these lines, the following chapters discuss the use of massively parallel sequencing to analyze prostate cancer genomes and exomes, building upon the studies reviewed here to expand our view of the genomic landscape of prostate cancer.

(Page intentionally left blank)

CHAPTER 2

Identification of driver mutations in prostate cancer using whole-exome sequencing

Adapted from:

*Barbieri, C.E., *Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N.*, et al.* (2012). Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. Nature Genetics *44*, 685-689.

* These authors contributed equally

Introduction

As described in the previous chapter, copy number alterations and oncogenic rearrangements are common in prostate cancer. These events include losses of *NKX3.1* (8p21) (Bhatia-Gaur et al., 1999; He et al., 1997) and *PTEN* (10q23) (Cairns et al., 1997a; Li, 1997), gains of the androgen receptor gene (*AR*) (Linja and Visakorpi, 2004; Visakorpi et al., 1995) and fusion of ETS-family transcription factor genes with androgen-responsive promoters (Perner et al., 2006; Tomlins et al., 2007; Tomlins et al., 2005). In contrast, recurrent somatic base-pair substitutions are believed to be less contributory in prostate tumorigenesis (Kumar et al., 2011; Taylor et al., 2010), but have not been systematically analyzed in large cohorts. To identify somatic mutations that may drive prostate cancer and aid in molecular characterization, we sequenced the exomes of 112 tumor/normal pairs enriched for characteristics of aggressive localized disease (pathological stage ≥ pT2, Gleason score ≥ 7). Novel recurrent mutations were identified in multiple genes, including *MED12* and *FOXA1*. *SPOP* was the most frequently mutated gene, with mutations involving the SPOP substrate binding cleft in 6-15% of tumors across multiple independent cohorts. *SPOP*-mutant prostate cancers lacked ETS rearrangements and exhibited a distinct pattern of genomic alterations. Thus, *SPOP* mutations may define a new molecular subtype of prostate cancer.

Whole exome sequencing of 112 primary prostate tumor–normal tissue pairs

We performed exome capture followed by paired-end, massively parallel sequencing on DNA from 112 prostate adenocarcinomas and matched normal samples. We focused on treatment-naïve radical prostatectomy specimens from American and Australian patients that spanned a range of grades, stages, and risk of recurrence (Table S2.1; please see Appendix 1: Supplemental tables). The exon

capture baits targeted 98.2% of genes in the Consensus CDS database as of November 2010 (http://www.ncbi.nlm.nih.gov/CCDS). A mean coverage depth of 118x per sample was achieved, with 89.2% of targets covered at ≥ 20x depth (Table S2.2 and Figure S2.1; please see Appendix 2: Supplemental figures). Tumor and normal DNA were also analyzed by Affymetrix SNP 6.0 arrays to detect somatic copy number alterations. In addition, transcriptome sequencing ("RNA-Seq") was performed on 22 exome-sequenced tumors and 41 independent samples (Figure S2.2).

We identified 5,764 somatic mutations that were present in tumor DNA but absent in peripheral blood or non-cancerous prostate (Table S2.3). Of these, 997 variants occurred in a single tumor that harbored a frame-shift mutation of the mismatch-repair gene *MSH6* (Figure S2.3). After excluding this highly-mutated sample, the remaining tumors contained a median of 10 silent and 30 non-silent mutations (range 10 to 105 total mutations) or ~1.4 per Mb covered (Figure S2.3). Analysis of 229 non-silent mutations by mass-spectrometric genotyping validated 95.6% of variants with allelic fraction ≥ 0.2 (C.I. 92-98%) (Table S2.3). The mutation rate of this cohort exceeded that of seven published prostate tumor genomes (0.9 mutations per Mb) (Berger et al., 2011), perhaps because the increased exome sequence coverage improved detection of variants present at lower allelic fractions.

Distinct mutational characteristics of tumor subclasses

We investigated whether pathological features corresponded to different mutational spectra. Pathologic stage pT3 tumors contained more mutations than pT2 tumors ($p = 1.2 \times 10^{-3}$, rank sum test) despite equivalent tumor purity between these classes (Figures S2.4 and S2.5). Substitutions in *PTEN* and *PIK3CA* were enriched in pT3 tumors ($p = 0.011$, Fisher's exact test) (Table S2.4), suggesting that these mutations may play a role in disease progression. Consistent with this possibility, activation of the PI3-Kinase pathway in mouse models accelerates the progression of prostate cancer (McMenamin et al., 1999; Trotman et al., 2003). This finding will need to be extended to larger panels of tumors due to the relatively small number of *PTEN* and *PIK3CA* mutations reported here. Interestingly, the base mutation

rate showed no correlation with Gleason score (a histological measure of disease risk) (Figure S2.5), indicating that mutational burden does not track uniformly with disease aggressiveness.

In addition, we noted that the mutational spectrum varied between prostate tumors harboring the *TMPRSS2-ERG* fusion and fusion-negative tumors. *TMPRSS2-ERG* fusion-positive tumors showed an increased proportion of CpG to T transitions ($p = 2 \times 10^{-4}$, Figure S2.5) but did not harbor more mutations overall. Since CpG to T transitions can arise from deamination of methylcytosine in cancer, this trend may reflect the differential methylation of DNA between ETS fusion-positive and fusion-negative tumors that was recently reported (Kim et al., 2011b) or may indicate a distinct mutagenic process in fusion-positive tumors.

Significantly mutated genes in prostate cancer

We searched for genes that harbored more non-synonymous mutations than expected by chance given gene size, sequence context and the frequency of mutations for each tumor (Figure 2.1A and Table S2.5). Twelve genes were enriched for mutations at q-value < 0.1, the majority of which are highly expressed at the transcript level in prostate tumors (Figure S2.6). The identification of *PIK3CA*, *TP53* and *PTEN* confirmed that our approach detected alterations known to promote tumorigenesis in prostate cancer and other malignancies. We also found evidence of enrichment for mutations in the *PTEN* pathway, cell cycle regulatory machinery, and other gene sets (Table S2.6).

The most frequently mutated gene was *SPOP* (13% of cases; Figure 2.1), which encodes the substrate-binding subunit of a Cullin-based E3 ubiquitin ligase (Nagai et al., 1997; Zhuang et al., 2009). Although isolated *SPOP* mutations have been reported in prostate cancer (Berger et al., 2011; Kan et al., 2010), this gene has not been found significantly mutated in any malignancy. Several novel genes not previously known to undergo somatic alteration in prostate cancer were enriched for mutations, including *FOXA1*, *MED12, THSD7B, SCN11A* and *ZNF595*. The $p27^{Kip1}$ gene *CDKN1B* was somatically mutated in three samples and deleted in sixteen others (Figure 2.1B). $p27^{Kip1}$ constrains prostate tumor growth in mice (Majumder et al., 2008) and harbors a germline prostate cancer risk allele (Kibel et al., 2003), but

Figure 2.1. Significantly mutated genes in aggressive primary prostate cancer

(**A**) (**Top**) A cohort of 111 primary prostate tumors is ordered by number of mutations per Mb sequenced. (**Center**) Mutations in significantly mutated genes, colored by the coding consequence of the mutation. Each column represents a tumor and each row a gene. (**Left**) Number and percentage of tumors with mutations in a given gene. (**Right**) The negative log of the q-values for the significance level of mutated genes is shown (for all genes with $q < 0.1$; see Methods).

(**B**) Net frequency of gene deletion/amplification across 169 copy number-profiled tumors. Significantly mutated genes are indicated. Only autosomal genes with two or more mutations are shown.

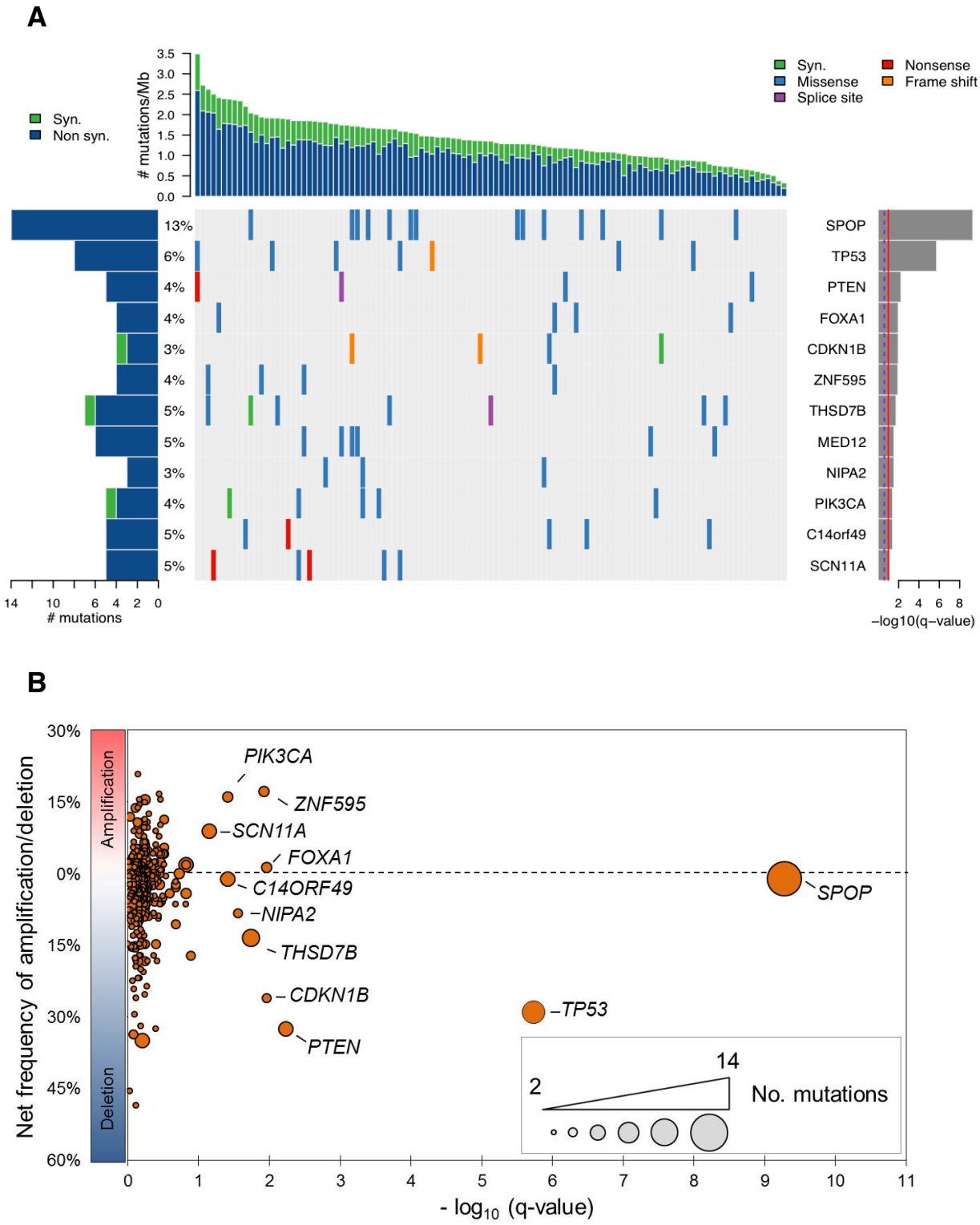


Figure 2.1 (continued)

somatic substitutions have not previously been observed in this cell cycle regulatory protein.

The Forkhead transcription factor gene *FOXA1* harbored nonsilent mutations in 4 of 111 exomes and 4 of 41 independent RNA-Seq samples. FOXA1 is required for epithelial cell differentiation in the murine prostate (Gao et al., 2005) and promotes cell cycle progression in castration-resistant prostate cancer (Zhang et al., 2011). Notably, FOXA1 modulates AR-driven transcription (Gao et al., 2003) and activates expression of *CDKN1B* (Williamson et al., 2006). Mutations strictly affected residues in the Forkhead domain that reside near the DNA binding surface (Figure 2.2A and Table S2.5) (Clark et al., 1993). The clustered nature of these mutations suggests that they may disrupt binding of FOXA1 DNA targets.

Mutations affecting *MED12* were observed in 6 out of 111 exomes, with a recurrent F1224L mutation in five samples (Figure 2.2B and Table S2.5). *MED12* encodes a subunit of the mediator complex and the Cyclin-dependent kinase 8 (CDK8) sub-complex that regulates basal and stimulus-specific transcriptional programs (Donner et al., 2007; Wang et al., 2002; Zhou et al., 2002). Recently, *MED12* mutations were reported in 70% of uterine leiomyomas (Makinen et al., 2011), benign stromal tumors of the smooth muscle lineage. Mutations in prostate cancer affected distinct codons from those in leiomyoma and occurred in epithelial cells rather than stroma as determined by laser-capture microdissection (LCM) (Supplementary Figure S2.7). Conceivably, *MED12* mutations may perturb CDK8-dependent modulation of transcriptional programs linked to p53 and androgen signaling (Donner et al., 2007; Wang et al., 2002).

Low-frequency mutations in cancer-associated genes

Multiple genes with established roles in other cancers were mutated at low frequency, including *IDH1*, *AKT1* and *HRAS* (Table S2.5). An analysis of predicted “damaging” mutations (nonsense substitutions, frame-shift indels and splice site alterations) in potential tumor suppressor genes expressed in prostate tumors identified mutations in *APC*, *PIK3R1* and *EPHA7* (Table S2.5). In addition, several chromatin-modifying enzymes harbored low-frequency damaging mutations, including *MLL1*, *MLL2,*

Figure 2.2 Recurrent somatic mutations in FOXA1 and MED12

(**A**) Structural analysis of mutations in FOXA1. Mutated residues are mapped to the structure of the HNF3γ fork-head domain from coordinate file 1VTN.pdb (www.pdb.org) (Clark et al., 1993) and highlighted in red. In both (A) and (B), mutations detected by exome sequencing are depicted (red), as are variants from non-overlapping transcriptome sequencing data (blue). FH, Fork-head domain.
(**B**) Recurrent MED12 mutations in prostate cancer (red, blue) are distinct from those reported in uterine leiomyeoma (shown in black) (Makinen et al., 2011). Domains of MED12 based on sequence content depicted based on Zhou et al. (Zhou et al., 2002). Multispecies conservation of the mutated sites is shown below the mutation.

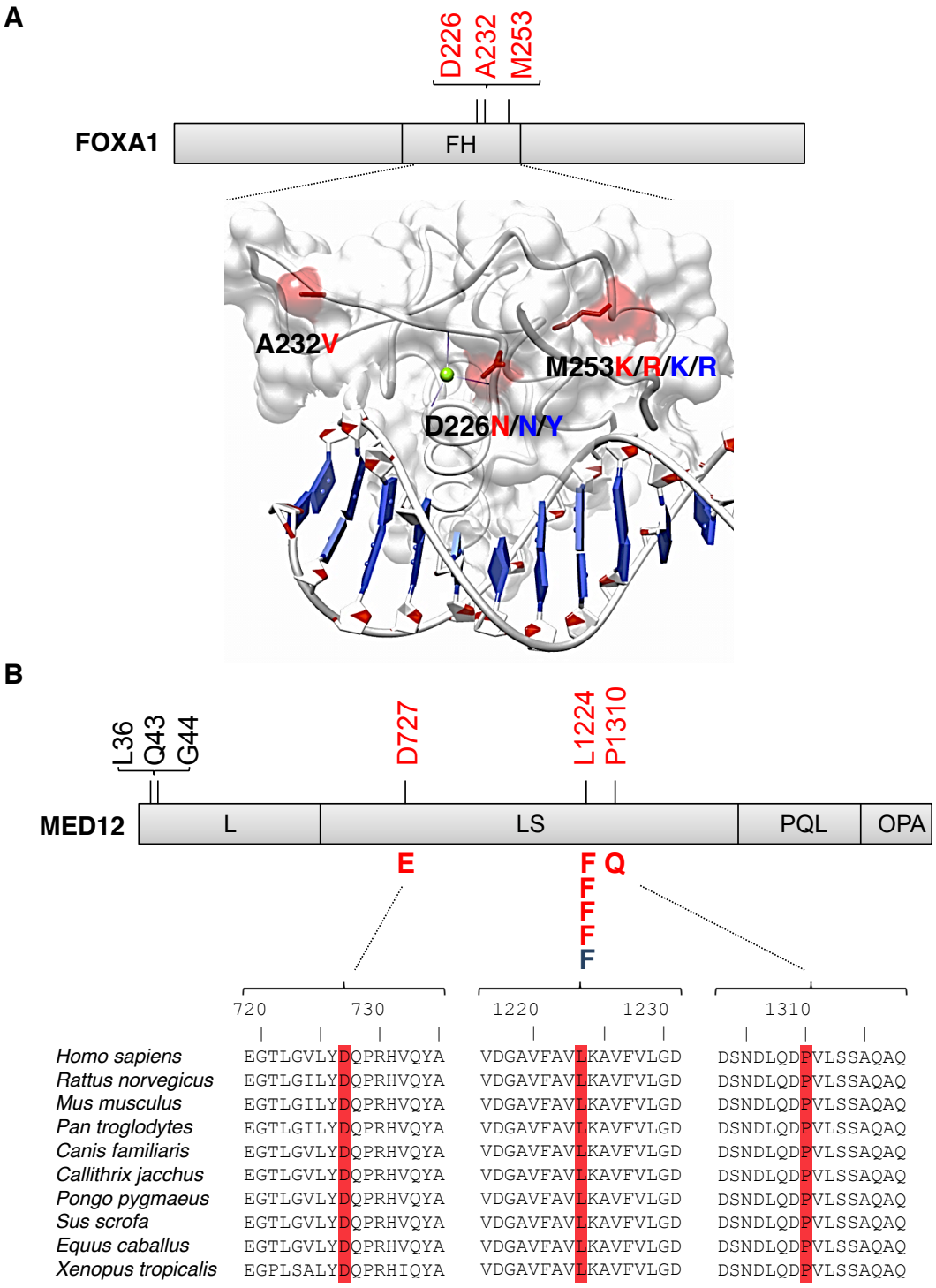


Figure 2.2 (continued)

MLL3, *ARID1A, NCOR1* and the histone demethylase gene *KDM6A* (*UTX*). Two *KDM6A* mutations involved residues situated within the catalytic Jumonji domain (I1209 and G1212), while a third introduced a frame-shift deletion directly N-terminal to this region (Table S2.5). These findings underscore the emerging importance of chromatin-modifying genes in prostate cancer (Gao and Alumkal, 2010; Grasso et al., 2012). Notably, *AR* was not mutated in any primary tumor analyzed, consistent with prior studies suggesting that mutations in this gene are restricted to metastatic or castration-resistant disease (Linja and Visakorpi, 2004; Taylor et al., 2010).

Recurrent mutations alter the substrate binding surface of the SPOP ubiquitin ligase complex protein

Although *SPOP* mutations were originally reported in genomic studies of prostate cancer (Berger et al., 2011; Kan et al., 2010), their prevalence and functional relevance remained unknown. We therefore sequenced this gene in multiple additional cohorts comprising over 300 primary tumors and metastases from the US and Europe. Using RNA-Seq and Sanger sequencing of tumor and matched germline DNA, recurrent heterozygous *SPOP* substitutions were identified in 6 to 13% of primary prostate adenocarcinomas (Figures 2.3A and S2.8, Table S2.7). No mutations were identified in 36 benign prostate tissue samples, prostate stroma, or 6 common prostate cell lines. *SPOP* mutations were also found in 6 of 41 patients with metastatic disease (14.5%). Thus, *SPOP* mutations occur at a 6 to 15% frequency across localized and advanced prostate tumors.

All SPOP mutations affected conserved residues in the structurally-defined substrate binding cleft (Figures 2.3B and Figure S2.9) (Zhuang et al., 2009). Several recurrently mutated residues exert key substrate-interacting roles; moreover, mutation of Y87, W131, and F133 disrupts substrate binding *in vitro* (Zhuang et al., 2009). These results strongly suggest that prostate cancer *SPOP* mutations are biologically consequential. To test this hypothesis, we examined the consequences of mutant SPOP protein expression or *SPOP* knock-down on tumorigenic phenotypes *in vitro*. Prostate cancer cells

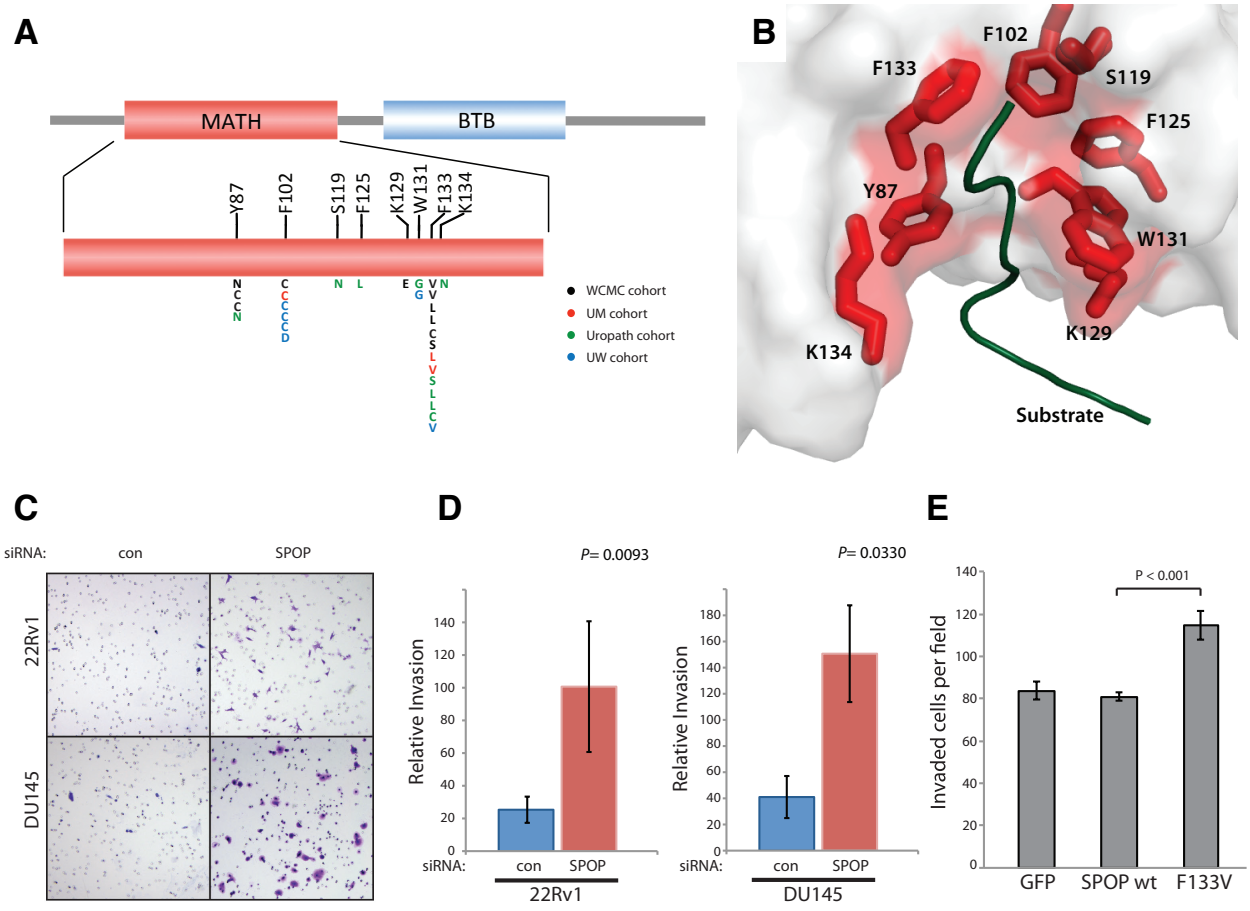


Figure 2.3. Structural and functional studies of recurrent SPOP mutations in prostate cancer

(**A**) Positional distribution of somatic mutations in SPOP across the Weill Cornell Medical College (WCMC), University of Michigan (UM), Uropath, and University of Washington (UW) prostate tumor cohorts.

(**B**) Mutated residues in the crystal structure of the SPOP MATH domain bound to substrate (PDB 3IVV).

(**C**) Representative images of invasive 22Rv1 and DU145 cells transfected with control and SPOP siRNA in Matrigel invasion assays.

(**D**) Quantitation of invaded cells transfected with SPOP siRNA.

(**E**) Quantitation of invaded DU145 cells transfected with GFP, SPOP wt, and SPOP F133V.

transfected with the most common *SPOP* mutant (F133V) or *SPOP* siRNA showed increased invasion compared to controls (Figures 2.3C-E and S2.10), but cell growth and viability were largely unaffected (Figure S2.11). The SPOP-CUL3 complex affects a variety of substrates that impact multiple pathways, including hedgehog, JNK, and steroid receptor signaling cascades (Li et al., 2011; Liu et al., 2009a; Wang et al., 2002). *SPOP* undergoes amplification in other malignancies and may be overexpressed in renal cell carcinoma (Liu et al., 2009a); however multiple prostate cancer cohorts showed no evidence of *SPOP* amplification or up-regulation (Figure 2.1B, Figure S2.12). Conceivably, prostate cancer-associated *SPOP* mutations exert *de novo* gain of function alterations (e.g., a distinct substrate profile), dominant negative effects, or more subtle alterations in substrate specificity. Further studies are necessary to determine the specific ubiquitin ligase functions and cellular pathways deregulated by *SPOP* mutation in prostate cancer.

SPOP mutations define a molecular subtype of ETS fusion-negative prostate cancer

Strikingly, all exomes with *SPOP* mutations lacked the *TMPRSS2-ERG* fusion or other ETS rearrangements (Figure 2.4, Figure S2.13), present in roughly 50% of prostate cancers (Mosquera et al., 2009; Tomlins et al., 2005). This mutually exclusive relationship between *SPOP* mutation and *ERG* rearrangement ($p < 0.001$, Fisher's exact test) was confirmed in evaluable samples across all five cohorts tested (Figure S2.13), and within two distinct foci from an individual prostate tumor (Figure S2.14). Thus, *SPOP* mutation and ETS fusions may represent early and divergent driver events in prostate carcinogenesis. *SPOP* mutations were identified in LCM-analyzed high-grade intraepithelial neoplasia (HG-PIN) adjacent to invasive adenocarcinoma, further strengthening the premise that *SPOP* mutation comprises an early event in prostate tumorigenesis (Figure S2.15).

In light of prior studies suggesting that prostate cancer may be classified by co-occurring genomic alterations (Demichelis et al., 2009; Lapointe et al., 2007; Taylor et al., 2010) we investigated whether *SPOP*-mutant tumors were enriched for other genomic lesions (Figure 2.4). Recurrent somatic deletions

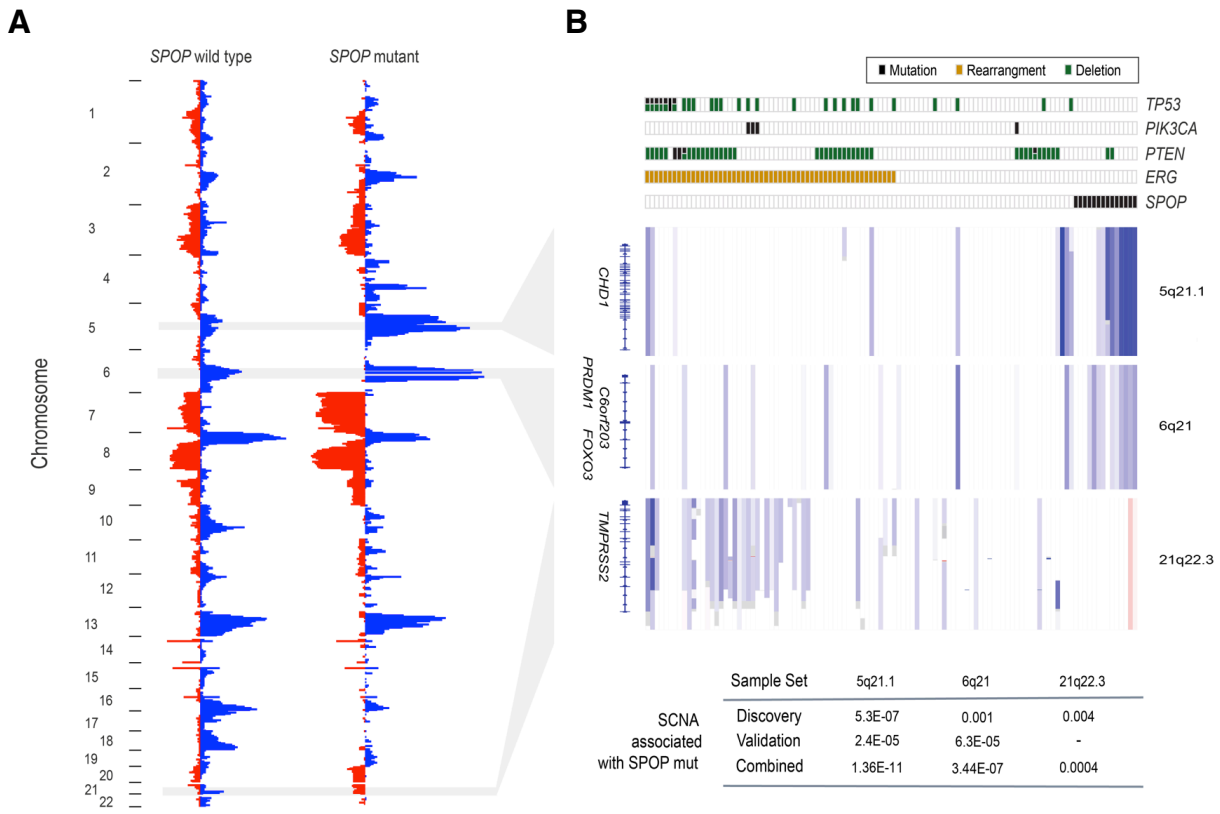


Figure 2.4. *SPOP* mutation defines a distinct genetic subclass of prostate cancer

(**A**) Frequency of genomic copy number alterations in *SPOP*-mutant and *SPOP*-wildtype tumors. Length of bars reflects the frequency of copy number loss (blue) or gain (red).

(**B**) Heatmap showing selected recurrent somatic copy number aberrations (SCNA). Each row represents a single prostate cancer sample. Samples are annotated for mutations in *SPOP*, *PTEN*, *PIK3CA*, and *TP53*, deletions of *PTEN*, and *ERG* rearrangements. Deletions positively correlated (5q21, 6q21) or inversely correlated (21q22.3) with *SPOP* mutation are shown. p-values of peak association with *SPOP* mutation in both discovery and validation cohorts are displayed at bottom (Fisher's exact test). Regions are not to scale; full coordinates available in Table S2.8.

at 5q21 and 6q21 were enriched in *SPOP*-mutant tumors (p = 1.4x10 $^{-11}$ and p = 3.4x10 $^{-7}$, respectively, Fisher's exact test) both in the whole-exome cohort and an independent prostate cancer collection (Figure 2.4, Table S2.8). Thus, loss of tumor-suppressor genes in these regions may collaborate with *SPOP* mutation to promote tumorigenesis. The relevant 5q21 locus contains *CHD1*, which encodes a chromatin-modifying enzyme that also undergoes disruptive rearrangements in prostate cancer (Berger et al., 2011). The 6q21 region encompasses several genes including *FOXO3*, a *FOXA1* homologue that has previously been implicated in prostate carcinogenesis and progression (Shukla et al., 2009), and *PRDM1*, a tumor suppressor in lymphoma (Mandelbaum et al., 2010). In contrast, *TP53* lesions were generally absent in *SPOP*-mutant tumors (p = 0.015, Fisher's exact test), despite the fact that this tumor suppressor gene was recurrently mutated and deleted (Figure 2.1). *SPOP* mutations also trended inversely with point mutations and/or copy number loss involving the *PTEN* locus in primary tumors (p = 0.044, Fisher's exact test) (Fig. 2.4); this pattern was supported by FISH analysis for *PTEN* deletion (Figure S2.16). *SPOP*-mutant tumors also lacked *PIK3CA* mutations (Figure 2.4). Although the inverse relationship between *SPOP* mutations and *PTEN/ PIK3CA* alterations was evident in primary tumors (p = 0.041, Fisher's exact test), these events co-occurred more frequently in metastatic tumors (Figure S2.16). Further studies are needed to determine if these genetic relationships also occur in other patient populations, and to elucidate the biological interactions that may underlie this phenomenon. Taken together, these results suggest that *SPOP* mutations may anchor a distinct genetic subtype of ETS-negative cancers.

Conclusion

In summary, whole-exome sequencing has identified genes that are recurrently mutated in prostate cancer. These efforts have also revealed a distinct ETS fusion-negative subclass of prostate cancer characterized by recurrent *SPOP* mutations and enriched for both 5q21 and 6q21 deletions. In the future, this expanded genetic framework may articulate new mechanisms of carcinogenesis that inform both disease modeling and patient stratification for clinical trials of experimental agents. Together with additional comprehensive analyses of the prostate cancer genome, epigenome, and transcriptome, these

systematic approaches should illuminate the landscape of alterations that underlie disease biology and therapeutic vulnerability in this common and clinically heterogeneous malignancy.

Acknowledgements

DNA sequencing was performed by the Broad Institute Genome Sequencing Platform. Christopher Barbieri performed the cellular invasion and proliferation experiments. Francesca Demichelis analyzed the RNA-Seq data. FISH and IHC assays were performed by Kyung Park, Naoki Kitabayashi and other members of the lab of Mark Rubin. Mutation significance analysis was performed in conjunction with Michael Lawrence.

Methods

Description of prostate tumor cohorts

Clinically localized primary prostate cancers were selected for exome- and transcriptome-sequencing from two cohorts: Weill Cornell Medical College (WCMC; New York, NY) and Uropath (Perth, Australia), a commercial supplier of banked urological tissues. Patients were included only if they had not received previous treatment for prostate cancer, including radiation therapy, brachytherapy or hormone ablation therapy.

Tumors from the WCMC cohort were collected by the Institutional Biobank from patients undergoing radical prostatectomy by one surgeon for clinically localized prostate cancer. Patient-matched normal DNA was obtained from whole blood samples as described below for this cohort.

Tumors from the Uropath cohort were obtained from men undergoing radical prostatectomy for clinically localized prostate cancer across multiple medical centers in Western Australia. Radical prostatectomies were performed by one of 30 clinicians between 2000 and 2010. Samples from both cohorts were stored at -80°C. Paired normal DNA was derived and sequenced from benign prostate

tissue. Normal DNA, was extracted from frozen tissue blocks with no histological evidence of neoplasia to minimize the possibility of contamination from tumor DNA.

For both cohorts, Hematoxylin and Eosin (H&E)-stained tissue sections were centrally reviewed by Juan-Miguel Mosquera, Kyung Park and Mark Rubin to verify Gleason score and to determine the percentage of Gleason pattern 4 and 5 histology at the site selected for DNA extraction. To characterize the ethnic composition of the cohorts, we analyzed high-density SNP array data by principal component analysis in combination with data from cohorts of known ethnicity from the HapMap database (CEU, YRI, CHB/JPT; http://hapmap.ncbi.nlm.nih.gov/) (Figure S2.17). All but five individuals chosen for exome sequencing clustered with CEU HapMap samples, indicating that patients were predominantly Caucasian. Four samples showed mixed or undetermined ethnicity and one clustered clearly with CHB/JPT (Han Chinese in Beijing/Japanese in Tokyo) HapMap samples.

In addition to exome- and transcriptome-sequenced tumors, prostate tumor cohorts from University of Michigan (UM), University of Washington (UW) and University Hospital Zurich (UZH) were used for extension screening for SPOP mutation. Prostate samples from the UM cohort were obtained from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program (Rubin et al., 2000), University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core (Ann Arbor, MI). Tumors from the UW cohort were obtained from the Rapid Autopsy Program, University of Washington and Fred Hutchison Cancer Research Center University (Seattle, WA). Samples from the UHZ cohort included a series of radical prostatectomy specimens, metastases, and benign prostatic hyperplasia samples. H&E–stained slides of all specimens were reevaluated by two experienced pathologists (Peter Wild and Colm Morrissey) to identify representative tissue sections. Tumor stage and Gleason score of the Zurich cohort were assigned according to the International Union Against Cancer and World Health Organization/International Society of Urological Pathology criteria.

All samples were collected with informed consent of the patients and prior approval of the institutional review boards (IRB) of respective institutions. Additionally, the sequencing and data release of all exome- and transcriptome- sequenced samples was reviewed and approved by local IRB.

DNA extraction for exome sequencing

H&E slides were cut from all frozen tissue blocks and examined by a board-certified pathologist to select for high-density cancer foci with <10% stroma or other noncancerous material to ensure high purity of cancer DNA. Biopsy cores were then taken from the corresponding frozen tissue block for DNA extraction. From each sample, 25-30mg of tissue was homogenized using a tissuelyser for 20 seconds at 15Hz. DNA was then isolated from the homogenate using the QIAamp mini kit (Qiagen) following the manufacturer's protocol. Samples were eluted in 150ul AE elution buffer and quantified using Picogreen dsDNA Quantitation Reagent (Invitrogen). Samples were qualified on an agarose gel (E-Gel, Invitrogen) to assess structural integrity. All DNA samples were stored at -20°C.

Whole exome capture library construction

Whole-exome hybrid capture libraries were constructed as described previously (Fisher et al., 2011), with minor modifications. Concentrations of tumor and normal DNA were measured using PicoGreen dsDNA Quantitation Reagent (Invitrogen). We sheared 100ng of genomic DNA to a range of 150-300bp using the Covaris E220 instrument. DNA fragments were end-repaired, phosphorylated, adenylated at the 3' termini and ligated to Illumina sequencing adapters as describe (Fisher et al., 2011), except that standard paired end adapters were replaced with forked adapters containing unique 8 base-pair index sequences (barcodes). Adapter-ligated DNA was then size-selected for lengths between 200-350bp and subjected to exonic hybrid capture using SureSelect v2 Exome bait (Agilent) according to (Fisher et al., 2011). The targeted exome covered 44Mb and comprised 98.2% of the CCDS database as of November 2010.

Library Quantitation and Sequencing

The number of properly adapter-ligated fragments in each library was quantified using quantitative PCR (qPCR) (Kapa Biosystems) with specific probes for the ends of the adapters. Based on qPCR quantification, libraries were normalized to 2nM and then denatured using 0.1 N NaOH. Barcoded

whole-exome libraries were pooled at equal molarities prior to sequencing with up to 93 samples per pool. Cluster amplification of denatured templates was performed according to manufacturer's protocol using V2 HiSeq Cluster Kits and V2 or V3 HiSeq Flowcells (Illumina). Paired end sequencing (2 x 76bp) was carried out on HiSeq Instruments, using V3 HiSeq Sequencing-by-Synthesis kits. The resulting data were analyzed with the current Illumina pipeline. Standard quality control metrics including error rates, % passing filter reads, and total Gb produced were used to characterize process performance prior to downstream analysis. The 8bp adapter index of each sequence read was used to match the read to its corresponding sample in the downstream data aggregation pipeline.

Exome sequence data processing

Two Broad Institute pipelines were used in succession to process and analyze exome sequencing data (Berger et al., 2011; Chapman et al., 2011; Stransky et al., 2011):

(1) The sequencing data processing pipeline "Picard", developed by the Sequencing Platform at the Broad Institute, starts with the reads and qualities produced by the Illumina software for all lanes and libraries and generates a BAM file (http://samtools.sourceforge.net/SAM1.pdf) representing each tumor and normal sample. The final BAM file stores all reads with well-calibrated qualities together with their alignments to the genome (only for reads that were successfully aligned).

(2) The Broad Cancer Genome Analysis pipeline, also known as "Firehose", starts with the BAM files for the tumor and matched normal samples and orchestrates various analyses, including quality control, local realignment, mutation calling, small insertion and deletion identification, coverage calculations and others (see details below).

Several of the tools used in these pipelines were developed jointly by the Broad Institute Sequencing Platform, Medical and Population Genetics Program, and Cancer Program as part of the Genome Analysis Toolkit (GATK) (http://www.broadinsitute.org/gatk). Additional details regarding these pipelines are provided elsewhere (DePristo et al., 2011; McKenna et al., 2010).

Sequence data processing pipeline (Picard)

For each sample, a BAM file was generated from Illumina sequence reads using the Picard pipeline (http://picard.sourceforge.net/) as previously described (Berger et al., 2011; Chapman et al., 2011). Briefly, Picard executes four steps: (1) alignment of sequence reads to the genome; (2) recalibration of base qualities based upon the quality score given by the Illumina software, the read-cycle, the lane, the tile and the identity of the base and the preceding base; (3) aggregation of lane-level and library-level data into a single BAM file per sample; and (4) marking of artifactual duplicate read pairs. These steps were performed as in (Chapman et al., 2011), with the following modification: sequence reads were aligned to the NCBI Human Reference Genome GRCh37 using the Burrows-Wheeler Aligner (BWA) (http://bio- bwa.sourceforge.net) (Li and Durbin, 2009).

The 224 BAM files produced by Picard, corresponding to 112 tumors and 112 normal samples, were deposited to dbGAP under accession no. phs000447.v1.p1.

Cancer genome analysis pipeline (Firehose)

We used Firehose, a cancer genome analysis pipeline infrastructure developed at the Broad institute, to analyze exome sequence data. The Firehose interface manages input files, output files and a variety of analysis tools. Firehose submits input files and parameters to GenePattern (Reich et al., 2006), which executes the specified modules or analyses. The analyses described below were performed as in (Berger et al., 2011; Chapman et al., 2011), with modifications where indicated.

Quality control

We used quality control modules in Firehose to ensure that each tumor and normal file corresponded to the correct individual, and that no mix-ups had occurred between tumor and normal data for a given individual. We compared genotypes from Affymetrix SNP 6.0 arrays and from tumor and

normal sequence data for each individual to ensure that all data corresponded to the correct patient. Genotypes from SNP arrays were also used to monitor for low levels of cross-contamination between samples from different individuals in sequencing data with the ContEst algorithm (Cibulskis et al., 2011).

Identification of somatic single nucleotide variants (SSNVs)

We used the MuTect algorithm from the Broad Institute Genome Analysis Toolkit (GATK) to identify SSNVs (https:// confluence.broadinstitute.org/display/CGATools/MuTect). As previously described (Berger et al., 2011; Stransky et al., 2011), MuTect identifies candidate SSNVs by performing a statistical analysis of the bases and their read qualities in the tumor and normal BAMs at the genomic locus under examination. Base-pairs were required to be covered by at least 14 reads in the tumor and 8 in the normal for mutation detection.

MuTect first filters out reads with low quality scores or excessive mismatches. A statistical analysis is then performed to identify somatic mutations using Bayesian classifiers for the tumor and normal sequences at a given locus:

$$LOD_T = log_{10}\left(\frac{P(\text{observed data in tumor} \mid \text{site is mutated})}{P(\text{observed data in tumor} \mid \text{site is reference})}\right)$$

$$LOD_N = log_{10}\left(\frac{P(\text{observed data in normal} \mid \text{site is reference})}{P(\text{observed data in normal} \mid \text{site is mutated})}\right)$$

Thresholds were chosen for each statistic to achieve a sufficiently low false positive rate. Several post-processing filters are applied to remove artifactual calls. For example, mutations are excluded that appear solely at the 5' or 3' end of reads or that are identified in panels of genomes from non-cancerous tissue. A subset of mutations was chosen for independent validation (see “Validation of selected mutations by mass spectrometry genotyping”). Mutations in known cancer-associated genes were reviewed manually using Integrative Genomics Viewer (IGV) (Robinson et al., 2011).

Local realignment and detection of indels

To improve detection of small insertions and deletions (indels), reads in tumor and paired normal were jointly realigned at genomic locations harboring putative indels by the local realignment module in the GATK (DePristo et al., 2011). Putative indels were then considered at sites that were well covered in tumor and normal where the indel-containing allele was supported by 8 or more reads or 30% of all reads from the locus. Next, these indel calls were filtered based on local alignment statistics around the putative event, including the average number of additional mismatches per indel-supporting read, average mismatch rate and base quality in a small window around the indel (The Cancer Genome Atlas Network, 2011).

Determination of mutation rates

We calculated rates of base mutations per Mb using the mutations detected (SSNVs and indels) and coverage statistics. Mutations were partitioned into categories based on their relative frequency such as (1) a C in a CpG dinucleotide mutated to T (CpG C to T transition), (2) all other Cs mutated to T (non-CpG C to T transition), (3) mutation of any C to G or A and (4) mutation of A to any other base. Disruptive mutations such as frame-shift indels and nonsense mutations were also considered separately. Because mutations may accumulate in benign-appearing tissue, we determined whether the use of benign prostate as the source of normal DNA affected our ability to distinguish somatic alterations from germline events. Tumors with matched normal prostate (n = 89) did not show different rates of mutation from tumors with blood matched normal (n = 22) (Figure S2.4), suggesting that the use of normal prostate did not prevent the detection of tumor-specific mutations.

Identification of significantly mutated genes

We applied the MutSig algorithm from the Broad Institute to identify genes that were significantly enriched for mutations as previously described (Berger et al., 2011; Chapman et al., 2011) with two modifications. First, at most one mutation per gene was considered from each sample. Second, the observed number of silent mutations was used as a guide to the local background mutation rate. Briefly,

MutSig identifies genes that harbor more mutations than expected by chance given sequence context and genic territory. Because certain base contexts exhibit increased mutation rates, such as cytosine in CpG dinucleotides, the context-specific mutation rates are considered for each class of mutation listed under "Determination of mutation rates" above. For each gene, we calculated the probability of detecting the observed constellation of mutations or a more extreme one, given the background mutation rates calculated across the dataset. This was done by convoluting a set of binomial distributions, as described previously (Getz et al., 2007). This p-value is then adjusted for multiple hypotheses according to the Benjamini-Hochberg procedure for controlling False Discovery Rate (FDR) (Benjamini, 1995) to obtain a q-value. The hyper-mutated sample (PR-00-1165) was excluded from this analysis.

Identification of significantly mutated gene sets

We also used MutSig to determine whether particular gene sets were enriched for mutations (Berger et al., 2011; Chapman et al., 2011). We downloaded the list of canonical pathways used in Gene Set Enrichment Analysis (GSEA) and analyzed 616 gene sets corresponding to known pathways or gene families. For each gene set, we tabulated the number of mutations occurring in any component gene, as well as the total number of covered bases in all genes in the gene set. A p-value was calculated for each gene set as for each gene, then a q–value was computed to account for the list of 616 hypotheses. The hyper-mutated sample (PR-00-1165) was excluded from this analysis. Indels called in significantly mutated genes and other cancer-associated genes were manually reviewed by inspecting the tumor and normal BAM files in IGV.

Mutation annotation

Point mutations and indels identified were annotated using Oncotator (Ramos *et al*, submitted) which integrates information from publicly available databases. In brief, a local database of annotations compiled from multiple public resources was used to map genomic variants to specific genes, transcripts, and other relevant features. The set of 73,671 reference transcripts used were derived from transcripts

from the UCSC Genome Browser's UCSC Genes track (Fujita et al., 2011) and microRNAs from miRBase release 15 (Kozomara and Griffiths-Jones, 2011) as provided in the TCGA General Annotation Files 1.0 library (https://wiki.nci.nih.gov/display/TCGA/RNASeq+Data+Format+Specification). Variants were also annotated with data from the following resources: dbSNP build 132 (Sherry et al., 2001), UCSC Genome Browser's ORegAnno track (Griffith et al., 2008), UniProt release 2011_03 (Uniprot Consortium, 2011) and COSMIC v51 (Forbes et al., 2011).

Validation of selected mutations by mass spectrometric genotyping

In order to validate detected mutations with an orthogonal genotyping method, we chose 240 non-silent mutations (231 SSNVs and 9 indels) across 48 T/N pairs to assay by mass spectrometric genotyping using the iPLEX platform (Sequenom). We targeted 74 mutations in significantly-mutated genes or gene sets with a q-value <0.1 and mutations reported in COSMIC. The remaining 166 non-silent mutations were chosen at random. Because the rate of validation using this technology falls significantly when the mutant allele is present at low allelic fraction (Berger et al., 2011; Stransky et al., 2011), we attempted to validate only mutations with AF ≥ 0.2 (i.e., where 20% of sequence reads from the tumor contain the mutation).

Of the 240 assays attempted, 228 gave successful genotype calls and 218 somatic mutations were confirmed (listed in Table S2.3). All events called in the tumor were absent from the corresponding normal. We conclude that the overall accuracy for mutation calling was 95.6% (CI: 92%-98%; Clopper-Pearson 95% confidence interval), in close agreement with previous studies (Berger et al., 2011; Chapman et al., 2011).

High-density SNP array analysis and detection of somatic copy number alteration

Genomic DNA from tumor and paired blood samples was processed using Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix, Inc.) according to manufacturer's protocols. The DNA was digested with NspI and StyI enzymes (New England Biolabs), ligated to the respective Affymetrix

adapters using T4 DNA ligase (New England Biolabs), amplified (Clontech), purified using magnetic beads (Agencourt), labeled, fragmented, and hybridized to the arrays. Following hybridization, the arrays were washed and stained with streptavidin- phycoerythrin (Invitrogen). Following array scanning, data preprocessing was performed using Affymetrix Power Tools. Copy number data was evaluated after segmenting the log2 ratios between tumor and paired normal levels on a sample basis. Quality control, data integrity, segmentation and copy number analysis were performed as previously described (Demichelis et al., 2009) with one additional step aimed at diminishing the number of recurrent lesions possibly caused by germline signal: we applied the same detection pipeline on the normal DNA samples alone. All peaks detected in both analyses were excluded from the recurrent somatic copy number aberration list. Cleared lesions with q-value < 0.1 were retained for association analysis with gene mutation status. Two-tail Fisher's Exact Test was applied for all association tests. Copy number profile images were generated with IGV (Robinson et al., 2011).

Assessment of tumor purity

Because prostate tumors may contain significant amounts of admixed stroma, we determined whether the purity of cancer DNA limited our ability to detect mutations. We assessed tumor purity by considering the allelic fractions (AF) of mutations detected in each tumor, defined as the number of mutation- supporting reads divided by the total number of reads mapping to a mutated locus. AF can be influenced by several factors in addition to tumor purity, including copy number alterations at mutated sites and the presence of subclones within a tumor. Therefore, we used AF data to assess purity in two ways. First, we considered the maximum mutant AF across all mutations in each tumor after removing the top fifth percentile of AF values in order to exclude outliers with elevated AF due to copy number alterations or stochastic effects. Second, we considered the median AF for all mutations across a tumor.

Both the maximum and median AF values correlated only slightly with the number of mutations detected, suggesting that tumor purity was not a systematic barrier to identifying mutations. In order to compare mutation rates across subgroups of tumors (e.g. Stage pT2 versus pT3), we verified that no

systematic differences in tumor purity existed between subgroups. To this end, we compared mutant and maximum AF for each subgroup and identified no differences (Figure S2.4).

RNA extraction, RNA-Seq sample preparation and sequencing

RNA was extracted from the frozen cancer tissue using TRIzol (Invitrogen) according to the manufacturer's protocol. Total RNA was prepared in accordance with Illumina's sample preparation protocol for paired end (PE) sequencing of mRNA as previously described (Pflueger et al., 2011). In brief, 5–10 µg of total RNA was fragmented by heat between 2 and 3 min based on the desired insert size, reverse-transcribed using Superscript II (Invitrogen), and transformed to double-stranded cDNA. To improve PE RNA-Seq data quality, we introduced an additional gel-based size selection step after cDNA double-strand synthesis and before the ligation of the PE adapters. This was postulated by Quail *et al.* (Quail et al., 2008) as a means to reduce the inclusion of artifactual chimeric transcripts that are composed of two cDNA fragments into the sequencing library. We also integrated the use of T4 ligase (Enzymatics Inc.) to improve the efficiency of adapter ligation. Over the course of the study, we increased the library size range from 250 bp to 450 bp. The gel dissolutions of all gel-based purification steps were conducted at room temperature under slight agitation as described by Quail *et al.* (Quail et al., 2008). After the enrichment of cDNA template by PCR, the concentrations and the sizes of the libraries were measured using a Qubit fluorometer (Invitrogen) and DNA 1000 kit (Agilent Technologies) on an Agilent 2100 Bioanalyzer, respectively. PE RNA- seq was performed with the Genome Analyzer II (Illumina) increasing the read size of the PE reads from 36 to 54 bp over the course of the study. Additionally, Illumina introduced improved sequencing reagents and upgraded imaging software over time to increase data quality and sequencing coverage.

Processing of RNA-Seq data

PE reads were aligned to the human genome (hg18) using ELAND, part of the standard software suite from Illumina, as previously described (Pflueger et al., 2011). Data were visualized using IGV (Robinson et al., 2011), and candidate mutations were identified in *SPOP* coding regions.

DNA extraction and SPOP genotyping

DNA was extracted using phenol-chloroform and purified by ethanol precipitation method as previously described (Berger et al., 2011). Direct Sanger sequencing of putative *SPOP* somatic mutations in all tumor-blood pairs was performed by standard methods following PCR amplification using specific primers. Sequences of the primers used for amplifying and sequencing *SPOP* are given in Table S2.9.

Laser-capture microdissection

5µm-thick tissue sections were cut, fixed and stained on membrane coated slides followed by dissection with the Arcturus*XT*™ LCM Instrument (Life Technologies Corporation, California, USA). Tissue staining and Laser Capture Mircodissection (LCM) were performed by Mirjam Blattner and Kyung Park as described by Espina *et al.* (Espina et al., 2006). A combined IR capture and UV laser cutting was carried out to best recover a precise subset of cells. DNA was amplified as suggested by manufacturer with the Whole Genome Amplification kit (WGA4) for single cell approach (Sigma Aldrich). Standard PCR was used for targeted enrichment of SPOP exon 6 and 7 followed by Sanger sequencing.

Fluorescence *in situ* hybridization

The *ETS* rearrangement status and *PTEN* deletion status was assessed on tissue slides from the same tumor nodule used for RNA and DNA extraction. Methods for fluorescence in situ hybridization (FISH) for *TMPRSS2-ETS* gene fusion have been previously described (Perner et al., 2006; Tomlins et al., 2005). We used *ERG, ETV1, ETV4,* and *ETV5* break-apart FISH assays to confirm gene rearrangement on the DNA level (Svensson et al., 2011). To assess the status of *PTEN*, we used a locus

specific probe and a reference probe as previously described (Berger et al., 2011). All FISH probes are listed in Table S2.10.

Quantitative RT-PCR

RNA was extracted using the TRIzol reagent (Invitrogen), subjected to DNase treatment (DNA-free kit; Applied Biosystems) according to the manufacturer's instructions, and used in quantitative RT-PCR. Quantitative RT-PCR was performed using the ABI 7500 Real-Time PCR System (Applied Biosystems) following the manufacturer's RNA-to-CT 1-step protocol. Each target was run in triplicate, and expression levels relative to the housekeeping gene GAPDH were determined on the basis of the comparative threshold cycle CT method ($2^{-\Delta\Delta}$CT). The primer sequences used in these experiments are given in Table S2.9. All experiments were run in triplicate; results are representative of three independent experiments.

Immunohistochemistry

ERG rearrangement status was confirmed by immunohistochemistry as previously described (Park et al., 2010). Briefly, primary rabbit monoclonal antibody was obtained from Epitomics. Antigen recovery was conducted using heat retrieval and CC1 standard, a high pH Tris/borate/EDTA buffer (VMSI). Slides were incubated with 1:100 of the ERG primary antibody for 1 hour at room temperature. Primary antibody was detected using the ChromoMap DAB detection kit and UltraMap anti-Rb HRP (VMSI). The anti-Rb HRP secondary antibody was applied for 16 minutes at room temperature. Slides were counterstained with Hematoxylin II for 8 minutes followed by Bluing Reagent (VMSI) for 4 minutes at 37°C. Subjective evaluation of ERG protein expression was scored as positive or negative by study pathologists (Kyung Park, Juan-Miguel Mosquera and Mark Rubin)

SPOP wild-type and mutant plasmids

Wild-type SPOP construct was obtained from Origene with C-terminal myc and FLAG tags in a mammalian expression vector. The SPOP-F133V construct was generated using the QuikChange II site-directed mutagenesis kit (Agilent). All plasmids were confirmed with Sanger sequencing, and protein expression was confirmed with Western blot using antibodies to SPOP, myc, and FLAG.

Cell culture and transfection

The human prostate cancer cell lines 22Rv1, and DU145 and the benign prostate epithelial cell line RWPE were obtained from the American Type Culture Collection. 22Rv1 and DU145Cells were maintained in RPMI 1640 (Invitrogen) supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin. RWPE cells were maintained in Keratinocyte-SFM (Invitrogen) supplemented with human recombinant Epidermal Growth Factor and Bovine Pituitary Extract (BPE).

For siRNA transfection, RWPE (2.5×10^5 per well), 22Rv1 (4×10^5 per well), and DU145 (2×10^5 per well), cells were seeded on 6-well tissue culture plates. The next day, cells were transfected with 100 nM SPOP or nontargeting (control) siRNAs (ON-TARGETplus; Thermo Scientific) using Dharmfect 2 reagent (Invitrogen) according to the manufacturer's instructions. For plasmid transfection, DU145 (4×10^5 per well), cells were seeded on 6-well tissue culture plates. The next day, cells were transfected with 4 ug of pCMV6-WT SPOP or pCMV-SPOP-F133V using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions.

Cell viability and proliferation assays

22Rv1 (2×10^3 per well) and DU145 (1×10^3 per well) cells transfected with control or SPOP siRNA or SPOP plasmids were seeded in 96-well tissue culture plates. Cell viability and growth was determined by performing WST-1 assay (Roche) reading absorbance at 450 nm according to the manufacturer's instructions. Values from three wells were obtained for each treatment and timepoint. Results are representative of three independent experiments.

Invasion assays

For invasion assays, 7.5×10^4 22Rv1 and 5×10^4 DU145 cells transfected with control or SPOP siRNA or SPOP plasmids were resuspended in 0.5 mL of RPMI-1640 medium containing 1% FBS and placed into the top chamber of Matrigel-coated 8-μm Transwell inserts (BD Falcon). The bottom wells contained RPMI supplemented with 5-10% FBS. After 24h (DU145) or 48h (22Rv1), the filters were fixed and stained with Crystal Violet 0.5% for 30 min, and cells on the upper surface of the filters were removed with a cotton swab. Migrated cells were quantified by counting the numbers of cells that penetrated the membrane in four microscopic fields (at 20X objective magnification) per filter. All experiments were run in triplicate; results are representative of three independent experiments.

(Page intentionally left blank)

CHAPTER 3

Complex DNA rearrangements result from punctuated genome-damaging events

Adapted from:

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M, Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., *et al.* (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666-77.

Introduction

Genetic studies of prostate cancer have revealed numerous recurrent DNA alterations that dysregulate genes involved in prostatic development, chromatin modification, cell cycle regulation and androgen signaling, among other processes (Baca and Garraway, 2012). Chromosomal deletions accumulate early in prostate carcinogenesis and commonly inactivate tumor suppressor genes (TSGs) such as *PTEN*, *TP53* and *CDKN1B* (Shen and Abate-Shen, 2010). In addition, exome sequencing of localized and castration-resistant prostate cancer has identified base-pair mutations in genes such as *SPOP, FOXA1* and *KDM6A*, which implicate a range of deregulated cellular processes in prostate tumor development (Barbieri et al., 2012; Grasso et al., 2012; Kumar et al., 2011).

Structural genomic rearrangements also play a critical role in prostate carcinogenesis. Roughly half of prostatic adenocarcinomas overexpress an oncogenic ETS transcription factor gene (most commonly *ERG*) due to somatic fusion with a constitutively active or androgen-regulated promoter (Tomlins et al., 2007; Tomlins et al., 2005). In addition, disruptive rearrangements may inactivate TSGs such as *PTEN* or *MAGI2* (Berger et al., 2011). Interestingly, analysis of prostate cancer genomes has revealed complex "chains" of rearrangements, which may result when broken DNA ends are shuffled and re-ligated to one another in a novel configuration (Berger et al., 2011). In theory, these DNA-shuffling events could simultaneously dysregulate multiple cancer genes, but the prevalence and consequences of rearrangement chains could not be assessed with the small panel of tumors sequenced.

Given the importance of structural genomic alterations in prostate cancer genesis and progression, we performed whole genome sequencing (WGS) and DNA copy number profiling of 57 prostate cancers to define a spectrum of oncogenic events that may operate during prostate tumor development. Through integrated analysis of rearrangements and copy number alterations, we discovered that the chromosomal disarray in a typical tumor may accumulate over a handful of discrete events during tumor development. We employ the term "chromoplexy" to describe this phenomenon of complex genome restructuring (from the Greek *pleko*, meaning to weave or to braid). Chromoplexy occurs in the majority of prostate cancers and may commonly inactivate multiple tumor-constraining

genes in concert, likely within a single clonogenic cell. This chapter describes the characterization of chromoplexy in prostate cancer genomes, and discusses how this knowledge informs a model for punctuated tumor evolution with relevance to prostate cancer and possibly other malignancies.

The landscape of genomic rearrangement in prostate cancer

We sequenced the genomes of 55 primary prostate adenocarcinomas and two neuroendocrine prostate cancer (NEPC) metastases that developed following castration-based therapy, along with paired normal tissue. We selected treatment-naïve adenocarcinomas across a range of clinically relevant tumor grades and stages (Gleason score 6 through 9; pathological stage pT2N0 through pT4N1; Table S3.1). Roughly 1.68×10^{13} sequenced base pairs aligned uniquely to the hg19 human reference genome (Table S3.2). Sequencing of tumor and normal DNA to mean coverage depths of 61x and 34x, respectively, revealed 356,136 somatic base-pair mutations, with an average of 33 non-silent exonic mutations per primary tumor (Figure 3.1 and Table S3.3A). We profiled somatic DNA copy number alterations (SCNAs) with high-density oligonucleotide arrays (Table S3.3B). Additionally, we conducted transcriptome sequencing on 20 tumors, along with matched benign prostate tissue for 16 cases.

To identify genomic rearrangements, we analyzed paired-end sequencing reads that map to the reference genome in unexpected orientations using the dRanger algorithm (Berger et al., 2011). We observed 5596 high-confidence rearrangements that were absent from normal DNA in both this cohort and an extended panel of 172 non-cancerous genome sequences (Figure 3.1 and Table S3.3C). We validated 113 rearrangements by re-sequencing and/or PCR amplification of tumor and normal DNA (Table S3C). We did not discover novel recurrent gene fusions, but observed several singleton events that may lead to overexpression of oncogenes. For example, sense-preserving fusions joined *NRF1* to *BRAF* (PR-4240) and *CRKL* to the ERK-2 kinase gene *MAPK1* (P04-1084; Figure S3.1A), leaving the kinase domains of *BRAF* and *MAPK1* primarily intact. Several genes underwent recurrent disruptive rearrangements with potential biological consequence, such as *PTEN*, *RB1*, *GSK3B* and *FOXO1* (Figure

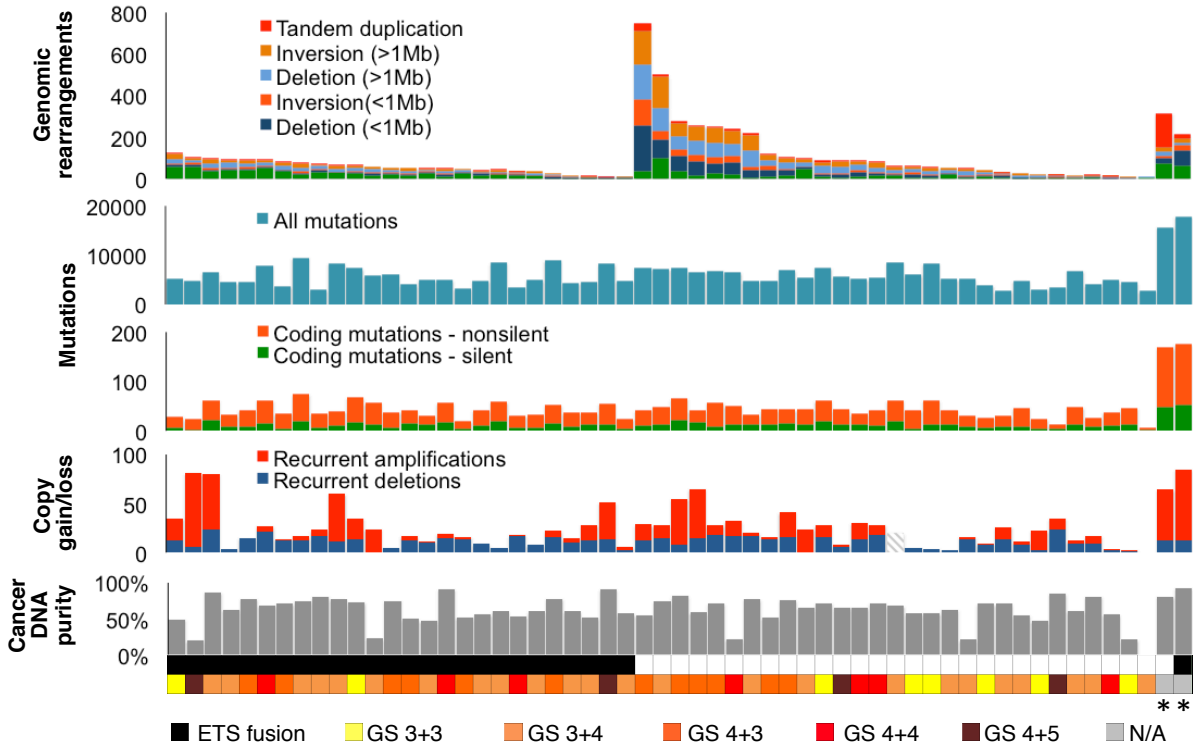


Figure 3.1. Somatic alterations in 57 prostate tumor genomes

WGS was conducted on 55 prostate adenocarcinomas and two lung metastases from neuroendocrine prostate cancers (NEPC, *) along with paired normal DNA to detect somatic rearrangements and mutations. Gains and losses of DNA copy number at sites of recurrent SNCAs were detected with Affymetrix SNP 6.0 arrays (recurrent SCNAs were not assessed for sample P07-144, hatched lines). Bottom, cancer DNA purity was evaluated by assessing allelic ratios from sequence reads covering heterozygous single-nucleotide polymorphisms at sites of chromosomal deletion (see Methods). ETS gene fusions (*ERG*, *ETV1*) were detected by sequencing and validated by fluorescence *in situ* hybridization (FISH).

S3.1 and Table S3.4). Thus, rearrangement of these genes may contribute to development of localized prostate cancer.

DNA deletions and rearrangements reveal signatures of complex genome-restructuring events

Rearrangements involving cancer gene loci often occurred in the context of a “chain”, in which the two rearrangement breakpoints map to the reference genome near breakpoints from other rearrangements (Figure 3.2A, left). Such characteristic breakpoint distributions were observed in our initial study of seven prostate cancer genomes (Berger et al., 2011) and appear to reflect collections of broken DNA ends that are shuffled and ligated to one another in an aberrant configuration. Given the involvement of prostate cancer genes in rearrangement chains, we set out to survey chained rearrangements systematically to clarify their prevalence, mechanistic underpinnings, and potential biological consequences.

We first determined whether additional chains could be identified by integrative analysis of chromosomal deletions and rearrangements. Although rearrangement chains may arise with minimal loss of genetic material, substantial DNA deletions were often evident at the fusion junctions of chained rearrangements (Figure 3.2A, right). When these deletions are overlaid with somatic rearrangement locations on the reference genome, the deletions create “bridges” that span the sequence between breakpoints from two different fusions (Figure 3.2A, bottom right). In all informative tumors in our cohort, the breakpoints at either end of a deletion were more often fused to novel partners rather than to each other (thus creating “deletion bridges”, rather than “simple deletions”; Figure 3.S2A). Importantly, this observation indicates that the many rearrangements demonstrating DNA loss near a breakpoint may be linked by deletion bridges to additional rearrangements in a chain.

We next considered whether rearrangements in a chain might arise independently of one another, for instance, at loci that are predisposed toward fusion due to DNA secondary structure or nuclear proximity (Burrow et al., 2010; De and Michor, 2011). To investigate this, we created a

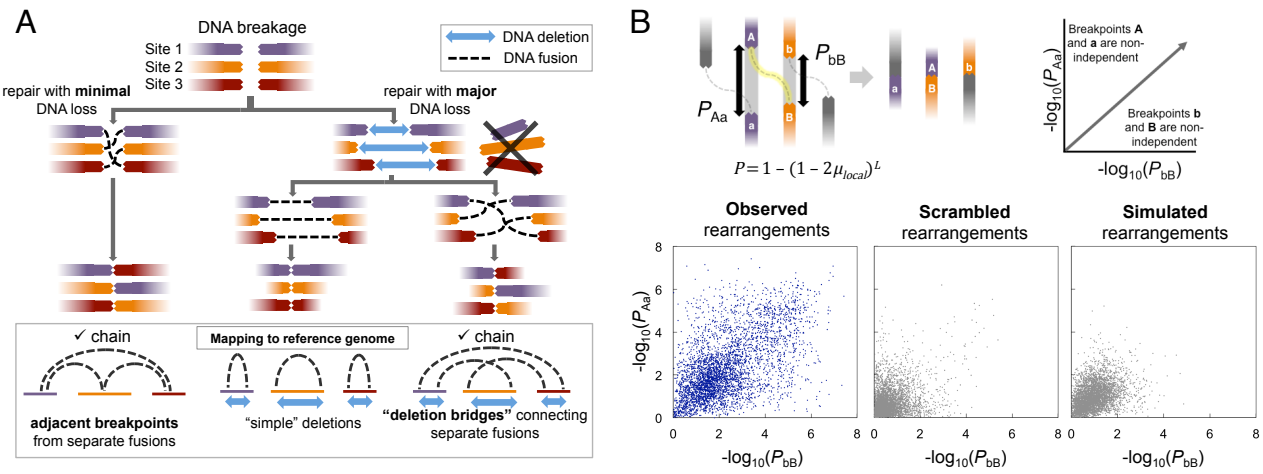


Figure 3.2. Integrated analysis of genomic deletions and rearrangements reveals signatures of inter-dependent alterations

(**A**) Three scenarios by which multiple DNA double-strand breaks may be repaired. Concerted repair with minimal loss of DNA (left) results in fusion breakpoints that map to adjacent positions in the reference genome. Loss of DNA at sites of double-strand breaks may result in simple deletions (middle) or "deletion bridges" (right) that span breakpoints from distinct fusions on the reference genome. Adjacent breakpoints or deletion bridges may provide evidence for chained rearrangements.

(**B**) For the two breakpoints of each rearrangement (labeled A and B), the probability *P* of a second independently generated breakpoint (a or b) falling within the observed distance (*L*) was assessed based on the expected local rate of rearrangements (μ_{local}). The x- and y- coordinates represent the negative log of *P* for the two breakpoints in each fusion. Rearrangements near the upper right corner of the plot are unlikely to have arisen independently of other rearrangements. Observed rearrangements are compared to simulated and scrambled data.

probabilistic model for the independent generation of detectible rearrangements across the genome (Figure S3.2B). Using this model, we calculated the probability that any pair of neighboring DNA breakpoints *X* and *Y* would arise independently of each other (P_{XY}) based on (1) their reference genome distance and (2) the local rate of rearrangements observed in our tumor panel (Figure 3.2B). As a control, we created ten simulated genomes for each tumor, with rearrangement locations matched for chromosome, local gene expression levels, sequence guanine/cytosine content and DNA replication timing, among other factors (see Methods). In addition, we generated "scrambled" genomes by drawing rearrangements from distinct tumors, preserving locus-specific effects that may promote double strand breakage. The observed rearrangements, but not the simulated or scrambled data, showed marked deviation from the independent model (Figure S3.2C) and statistical enhancement of chain-like patterns (Figure 3.2B). For 50% of rearrangements, the reference genome locations of both breakpoints were nearer to breakpoints of additional rearrangements than would be expected by chance ($p < 10^{-4}$ for observed versus simulated or scrambled P_{XY} values). To the extent that our model correctly predicts the genomic distribution of independent rearrangements, these results suggest that rearrangement chains are unlikely to arise from independent events, thus raising the hypothesis that they occur by a coordinated process.

"Chromoplexy" generates chained chromosomal rearrangements and deletions

Having identified chained patterns of adjacent rearrangements that may signify concerted alterations, we created an algorithm called ChainFinder to search for co-arising structural alterations (Figures 3.3A and S3.3). ChainFinder employs a statistically based search rooted in graph theory to identify genomic rearrangements and associated deletions that deviate significantly from our independent model described above, and thus appear to have arisen in an interdependent fashion (see Methods).

We used ChainFinder to survey our panel of prostate tumors for rearrangement chains. Strikingly, this analysis revealed numerous chains involving widely variable numbers of rearrangements. Some

Figure 3.3 The ChainFinder algorithm

(**A**) ChainFinder creates a graph representation of genomic breakpoints that may be linked in chains by somatic fusions, statistical adjacency or deletion bridges. ChainFinder assigns two neighboring breakpoints to the same chain if the p-value for their independent generation (*P*) is rejected with a false-discovery rate below 10^{-2}. For each cycle (closed path) within the graph, all scenarios are considered where one or more rearrangements in the cycle could have arisen independently. All rearrangements in a cycle are assigned to the same chain if every such scenario is rejected with a family-wise error rate below 10^{-2} across all scenarios. Please see the Methods for additional details.

(**B**) Circos plot of chained rearrangements in a prostate adenocarcima (P09-1042). Rearrangements depicted in the same color arose within the same chain; fusions in gray were not assigned to a chain. The inner ring depicts copy number gains and losses in blue and red, respectively.

(**C**) The false positive rate of ChainFinder was assessed using simulated and scrambled genomes based on observed rearrangements.

(**D**) For observed, simulated and scrambled genomes, the longest chain was compared along with the portion of breakpoints in any chain.

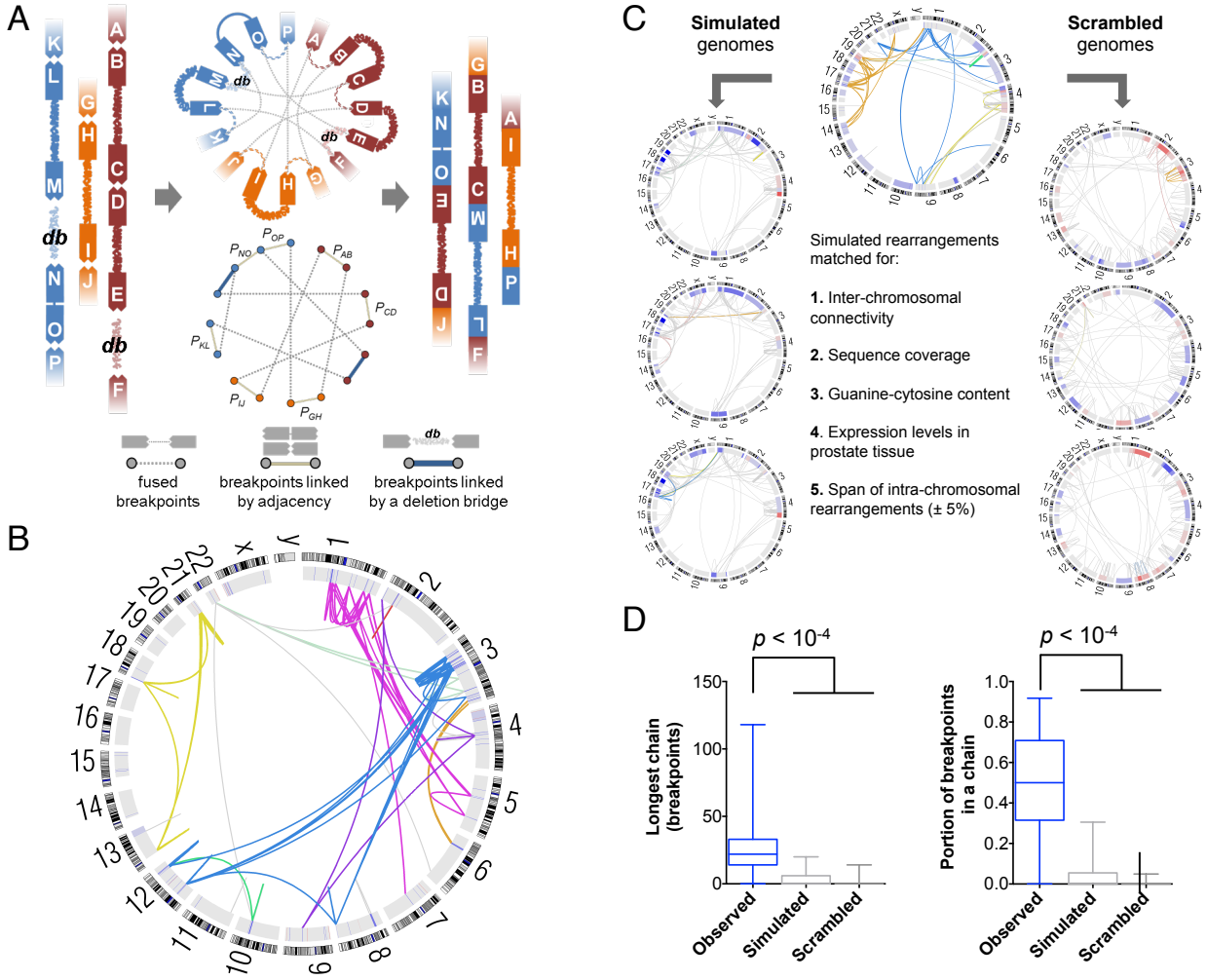


Figure 3.3 (continued)

chains involved only three fusions, while others revealed more than forty rearrangements that wove together five or more chromosomes (Table S3.5A; Figure 3.3B and S3.3C). We have termed the process of genomic restructuring that produces these complex chains “chromoplexy” (from the Greek *pleko*, meaning “to weave” or “to braid”). Chromoplexy-associated chains of five or more rearrangements (ten or more breakpoints) were detected in 50 out of 57 tumors (88%; Table S3.5B and Figure S3.3C), while 36 tumors (63%) contained two or more such chains. Overall, 39% of rearrangements participated in chains, while ChainFinder detected chains in only 2.8% and 0.2% of rearrangements from simulated or scrambled genomes, respectively (Figure 3.3C-D). Thus, chromoplexy generates multiple structural alterations in a coordinated fashion, as inferred by statistical analysis of breakpoint distributions.

We noted profound phenotypic differences in chromoplexy in subsets of prostate cancers. Chromoplexy in tumors harboring oncogenic ETS fusions (ETS^+) produced significantly more inter-chromosomal rearrangements than ETS^- tumors ($p < 10^{-4}$) and involved a greater maximum number of chromosomes in a single event ($p = 9 \times 10^{-3}$; Figure 3.4A-C). Interestingly, oncogenic *ERG* fusions frequently arose in the setting of chromoplexy (15 of 26 cases, 58%). Given that fusion of *TMPRSS2* and *ERG* occurs in the setting of androgen receptor-driven transcription (Haffner et al., 2010), the intricate chains in ETS^+ tumors could reflect DNA injury at transcriptional hubs occupied by loci from multiple chromosomes. Consistent with this possibility, chromoplexy in ETS^+ nuclei primarily affected regions of the genome that are highly expressed in prostate tumors (Figure 3.4D) and that co-localize in interphase nuclei (Figure S3.4A). Thus, chromoplexy in ETS^+ tumors appears to reflect a distinct process of genome restructuring that may be coupled to transcriptional processes.

In contrast, chromoplexy in a subset of ETS-negative cancers resembled chromothripsis (Rausch et al., 2012; Stephens et al., 2011), a process of chromatin shattering yielding extensive DNA rearrangement, often of one or two focal chromosomal regions. In particular, seven ETS^- tumors contained up to seven-fold more rearrangements than the whole-cohort average (Figure S3.4B). These tumors harbored focal deletions or disruptive rearrangements involving the chromatin-modifying enzyme gene *CHD1*, a putative tumor-suppressor gene that may regulate genomic stability (Huang et al., 2011; Liu et al., 2012). The rearrangements in $CHD1^{del}$ tumors were predominantly intra-chromosomal both within chains ($p = 2 \times 10^{-4}$) and overall ($p = 4 \times 10^{-4}$; Figure S3.4C). Moreover, the rearrangements in

Figure 3.4. Manifestations of chromoplexy vary by ETS fusion status

(**A**) Circos plots of rearrangement chains in representative tumors, grouped by the presence of ETS rearrangements and *CHD1* disruption. Rearrangements in the same chain are depicted in one color. Rearrangements in gray were not assigned to a chain. The inner ring shows copy number gain and loss in red and blue, respectively.

(**B**) Rearrangement chains in ETS-positive tumors contain a greater proportion of inter-chromosomal fusions than chains in ETS-negative tumors.

(**C**) The maximum number of chromosomes involved in a single rearrangement chain (y-axis), grouped by ETS status. The total number of breakpoints in chains in each tumor is depicted on the x-axis to allow comparison of tumors with similar degree of detectable chromoplexy.

(**D**) ETS-positive chromoplexy breakpoints are enriched near DNA that is highly expressed in 16 prostate tumor transcriptomes.

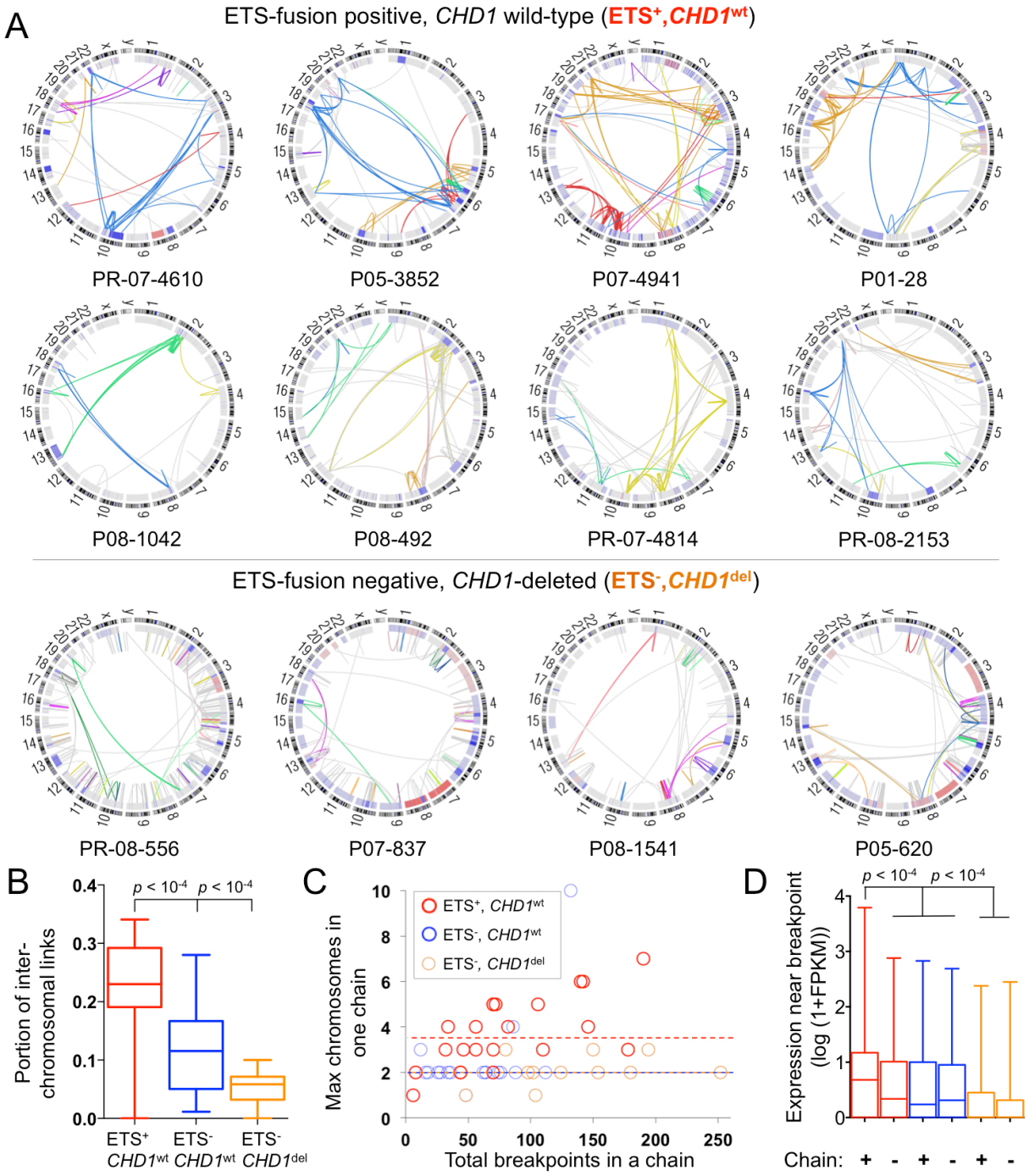


Figure 3.4 (continued)

CHD1^{del} samples arose in late replicating DNA with low guanine and cytosine content (Figure S3.4B), generally corresponding to gene-poor heterochromatin. An extended cohort of 199 prostate adenocarcinomas revealed that *CHD1* loss was associated with an increased number of recurrent SCNAs (p = 1.5 x 10^{-8}; Figure S3.4C). Given the postulated roles of CHD1 in genome stability and maintenance of chromatin architecture (Gaspar-Maia et al., 2009), these findings raise the possibility that *CHD1* deletion may contribute to the distinctive patterns of genomic instability observed in *CHD1*^{del} tumors.

We investigated whether chromoplexy is unique to prostate cancer by analyzing a panel of 59 additional tumor genomes including melanoma, non-small cell lung cancer, head and neck squamous cell carcinoma, and breast adenocarcinoma (Table S3.5B and Figure S3.3C). Every tumor type demonstrated multiple instances of chains involving 5 or more rearrangements. Thus, a small number of chromoplexy events may account for the wide array of rearrangements and deletions in several common cancers.

Chromoplexy commonly dysregulates cancer genes

To assess the role of chromoplexy in prostate cancer development, we examined the genomic regions altered by deletion or disruptive rearrangements in the context of chains. Using a list of 17 potential prostate tumor suppressor genes from the KEGG database (Kanehisa et al., 2012), we found that 26 of the 57 tumors (46%) have either deletion or rearrangement of at least one gene in a chain of three or more rearrangements (Table S3.5C). Inclusion of the *TMPRSS2-ERG* fusion and 10 putative prostate cancer genes added 9 more samples. Several cancer genes were recurrently deleted or rearranged by chromoplexy, including *PTEN* (9 cases), *NKX3-1* (8 cases), *CDKN1B* (3 cases), *TP53* (4 cases), and *RB1* (2 cases). Thus, chromoplexy may drive prostate carcinogenesis by disrupting tumor suppressor genes and creating oncogenic fusions.

The concurrent shuffling and deletion of multiple regions across the genome that may underlie chromoplexy could simultaneously inactivate tumor suppressor genes that are geographically distant from each other (i.e. on separate chromosomes). We noted several examples where multiple putative cancer

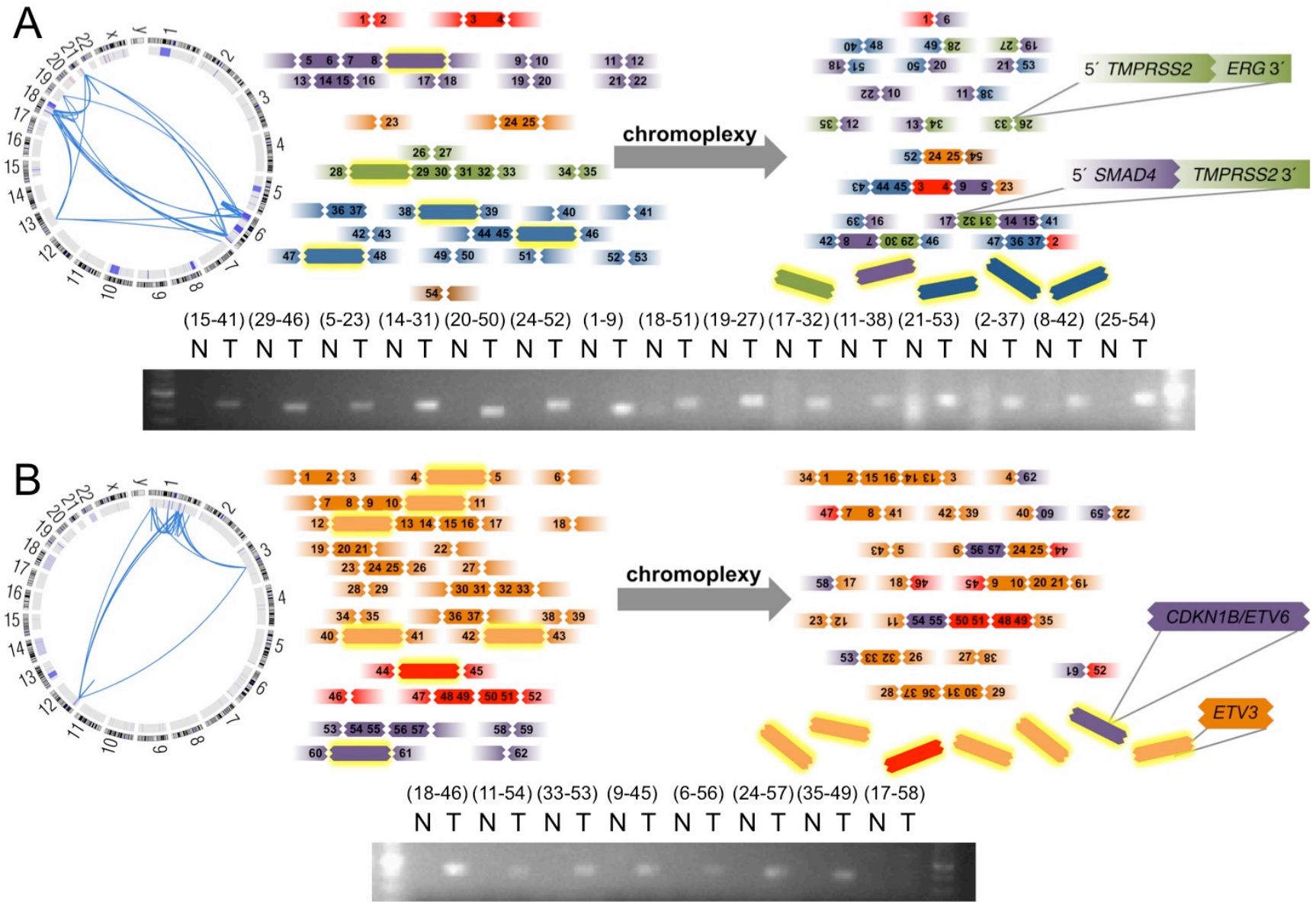
Figure 3.5. Chromoplexy generates widespread genomic alterations that coordinately dysregulate multiple cancer genes

(**A**) Chromoplexy-associated chain of 27 somatic rearrangements across 6 chromosomes in tumor P05-3852, involving fusion of *TMPRSS2* and *ERG* and disruptive rearrangement of *SMAD4.*

(**B**) The putative tumor suppressor genes *CDKN1B*, *ETV6* and *ETV3* were lost in the context of deletion bridges in a 25-rearrangement chain affecting 3 chromosomes in PR-05-3595.

In both panels, selected rearrangements were assessed by PCR of tumor and normal DNA.

Figure 3.5 (continued)



genes were disrupted by a single instance of chromoplexy. For instance, a chain of 27 rearrangements across 6 chromosomes coordinately fused *TMPRSS2* and *ERG* (21q) and disrupted the *SMAD4* (18q) prostate tumor suppressor gene (Ding et al., 2011) (Figures 3.5A and S3.5). In a second example, the adjacent *CDKN1B/ETV6* tumor suppressor genes (12p) and the *ETV3* locus (1q) were lost in the context of deletion bridges within one chain (Figure 3.5B). Additional instances of chromoplexy disrupted interacting genes in the same pathway: for instance, deletion of *PIK3R1* (5q) with *PTEN* (10q) and *TP53* (17p) with *CHEK2* (22q) occurred in two chains (Table S3.5C). Thus, chromoplexy may simultaneously dysregulate multiple cancer genes across the genome. Such events may provide selective advantages to incipient cancer cells, particularly given that the loss of some TSGs promotes prostate cancer only in the context of specific accompanying molecular lesions (Chen et al., 2005).

Clonal evolution reveals paths of prostate cancer progression

To provide additional insight into the genomic evolution of prostate tumors, we analyzed the clonal status of mutations and deletions in our cohort. Using an approach related to previously described methods (Carter et al., 2012; Nik-Zainal et al., 2012), we exploited the extensive germline SNP genotype data provided by WGS to assess tumor purity and the clonal status of genomic lesions (Figures 3.6A and S3.6). Our estimates of tumor purity based on WGS matched those produced by ABSOLUTE analysis of SNP array data (Carter et al., 2012) ($R^2 = 0.99$; $p < 10^{-4}$), with the exception of two samples where admixed normal DNA was detected only from WGS data (Table S3.1).

We first compared the clonality of deletions involving prostate cancer genes, reasoning that lesions that arise early in tumorigenesis or that foster rapid outgrowth would tend to be clonal, while late-arising deletions would more often be subclonal. Several common deletions were strictly clonal, including *NKX3-1* and the 3Mb region of chromosome 21q that is frequently deleted to produce the *TMPRSS2-ERG* fusion (Perner et al., 2006) (Figure 3.6B and Figure S3.6). These events are among the earliest detectible alterations in prostate cancer and are frequently observed in prostatic intraepithelial neoplasia (PIN), a prostate cancer precursor lesion (Emmert-Buck et al., 1995; Perner et al., 2007). By contrast,

Figure 3.6. Clonality and evolution of prostate cancer

(**A**) Schematic representation of the clonality assessment. The allelic fractions (AFs) of sequencing reads covering heterozygous SNPs were analyzed in order to assess the clonality of somatic DNA alterations. A hypothetical tumor is shown, composed of normal cells, a caner clone and a derivative subclone. The histograms indicate the expected SNP AFs within two deleted genes, A and B. The subclonal deletion of B yields a distinct distribution of AFs compared to the clonal deletion of A.

(**B**) Selected deletions (top) and mutations (bottom) were classified as clonal or subclonal. Proportion test p-value is listed for the indicated comparisons. Independent samples (Barbieri et al., 2012) are included for support.

(**C**) Example of clonal (*TMPRSS2-ERG*) and subclonal (*CDKN1B*) deletions from the same tumor. Histograms show the proportion of sequencing reads containing the reference allele for heterozygous SNPs in the deleted regions. A representative immunohistochemical stain for the *CDKN1B* protein $p27^{Kip1}$ shows discrete subclonal positivity in prostate cancer.

(**D**) Patterns of tumor evolution were inferred based on clonality estimates. Arrows indicate the direction of clonal-subclonal hierarchy between genes that are deleted in the same sample in multiple cases. Deleted genes are represented by circles with size and color intensity reflecting the frequency of overall deletions and subclonal deletions, respectively. Ratios along the arrows indicate the number of samples demonstrating directionality of the hierarchy out of samples with deletion of both genes (ratios in parentheses refer to additional samples; Barbieri et al., 2012). The inset shows a similar analysis of point mutations (Barbieri et al., 2012).

(**E**) The number of recurrent SCNAs and cancer DNA purity were compared across tumors with major Gleason pattern 4 versus 3.

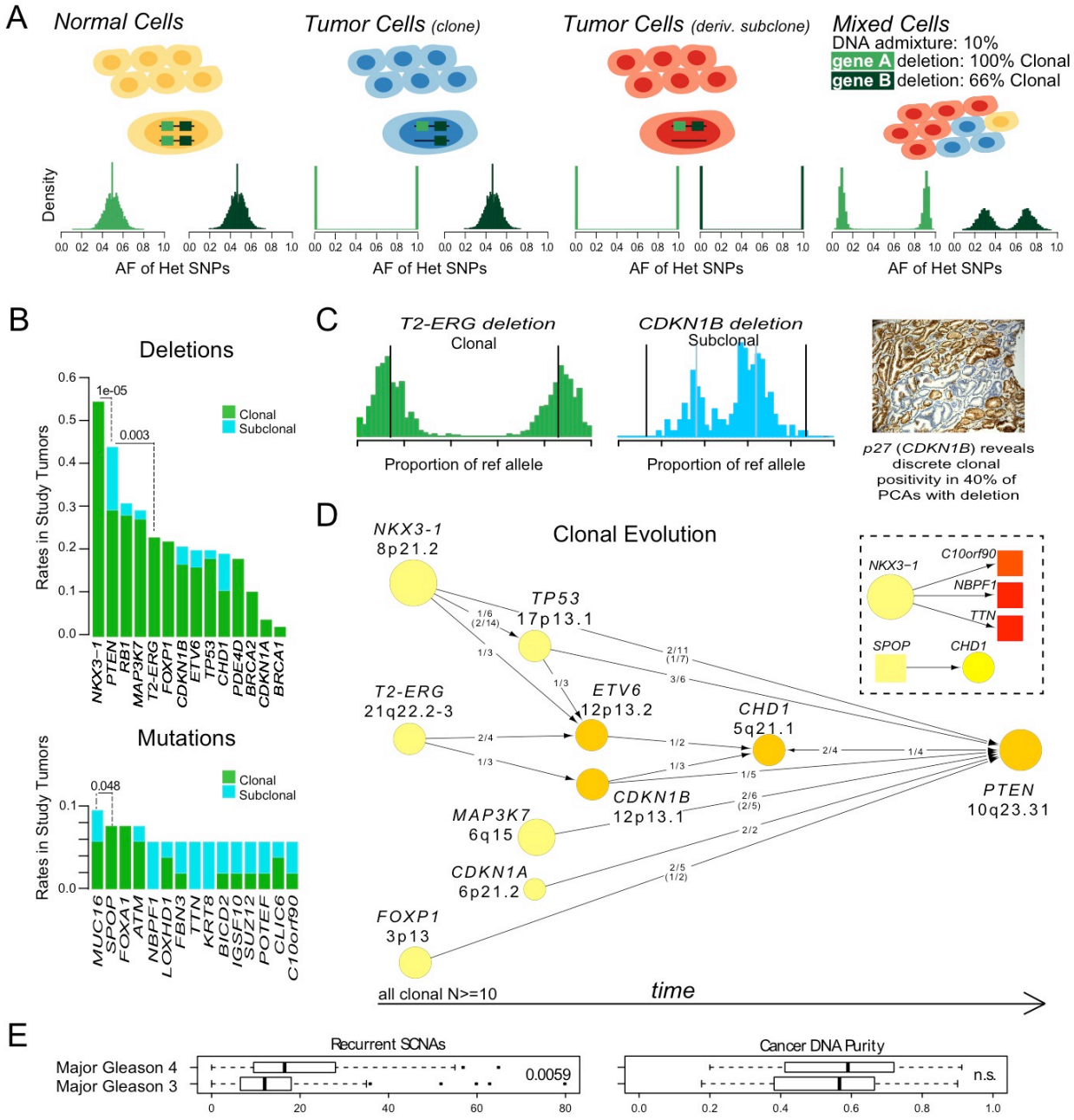


Figure 3.6 (continued)

deletions of *PTEN* were often subclonal ($p = 10^{-5}$ for comparison with *NKX3-1* deletion clonality), as were *CDKN1B* deletions (Figure 3.6C). This finding suggests that *PTEN* and *CDKN1B* inactivation promotes the early progression of prostate cancer, consistent with the association of these events with higher-stage disease (Barbieri et al., 2012; Halvorsen et al., 2003).

We next used our clonality assessments to deconvolve the sequence of oncogenic events that gives rise to a typical prostate tumor. Reasoning that clonal alterations must originate prior to subclonal alterations within the same tumor, we examined pairs of genes that were deleted in the same sample across multiple tumors to determine the directionality of the clonal-subclonal hierarchy (Figure 3.6D). Where possible, we confirmed these relationships in independent exome-sequenced tumors. A "consensus path" of progression emerged, beginning with events including deletion of *NKX3-1* or *FOXP1* and fusion of *TMPRSS2* and *ERG*. These lesions may disrupt normal prostate epithelial differentiation (Bhatia-Gaur et al., 1999; Sun et al., 2008) and effect other oncogenic perturbations. Thereafter, lesions in *CDKN1B* or *TP53* accumulate; these alterations may lead to enhanced proliferation, genomic instability and/or evasion of apoptosis. Finally, loss of *PTEN* may provide a gating event in the development of aggressive prostate cancers. A similar assessment of point mutation clonality (Figure 3.6B, lower) revealed higher overall rates of subclonal events, with the exception of early mutations as in *SPOP* and *FOXA1*. Together, these results imply that prostate carcinogenesis favors the dysregulation of cancer genes in defined sequences, as has been suggested by studies of developing tumors in colon cancer (Fearon and Vogelstein, 1990).

Next, we investigated whether chromoplexy might continue after cancer initiation, and thereby contribute to the progression of a tumor down an oncogenic path. Interestingly, several chains appeared to involve strictly subclonal deletion bridges (Figure S3.7A), indicating that tumors may sustain multiple rounds of chromoplexy. Together with the observation that chromoplexy targets both early and late genes in the consensus path (e.g., *ERG* and *PTEN*) these findings suggest that chromoplexy continues to drive the outgrowth of tumor subclones.

Prostate cancer genomic derangement increases with histological grade

Finally, we considered whether tumors with high-grade histology (indicative of high clinical risk) might occupy positions further along the consensus path. To this end, we quantified recurrent SCNAs in each genome by counting amplifications and deletions that overlapped with regions of significant SNCAs identified by GISTICv2 analysis (e.g., the *TP53*, *PTEN* and *CDKN1B* loci) across 199 tumors reported here and in a previous study (Barbieri et al., 2012; Beroukhim et al., 2010). Tumors with predominantly Gleason score (GS) 4 histology were significantly enriched for recurrent SCNAs compared to GS 3 tumors ($p = 5.9 \times 10^{-3}$; Figure 3.6E) beyond the overall extent of SCNAs, despite similar purity of cancer DNA and mutational burden between the two groups. Altogether, these findings suggest that structural alterations affecting cancer genes, many of which result from chromoplexy, may contribute to the aggressive clinical behavior of high-grade prostate tumors.

Discussion

We have characterized somatic alterations across the genomes of 57 prostate tumors. By systematically profiling rearrangements and copy number alterations, we identified chromoplexy as a common process by which multiple geographically-distant genomic regions may be disrupted at once. Chromoplexy is evident in several solid tumor types and in the majority of prostate cancers. In multiple instances, chromoplexy altered more than one cancer gene coordinately. In the future, systematic assessment of chromoplexy from WGS data could reveal groups of cancer gene alterations that confer a selective advantage when sustained all at once, but activate tumor-suppressing safeguards if sustained individually.

Although chained rearrangements could theoretically arise over multiple cellular generations by a "sequential-dependent" mechanism, where the occurrence of each subsequent event depends on the presence of a prior event (Figure S3.7B), such a mechanism seems unlikely. In particular, a sequential-dependent model fails to account for the many complete or "closed" chains we detected. For a closed

chain to arise in a sequential-dependent manner, multiple junctions from ancestral somatic fusions would have to be re-broken precisely and fused to each other (Figure S3.7B) to complete the chain. Even if breakpoints in a chain could only fuse to one another, to generate the 121 observed closed chains in a sequential-dependent process would require immensely elevated rates of rearrangement in a focused region of the genome (up to ~10^3 times the maximum observed rate; Figure S3.7C-D). While we cannot exclude this possibility, plausible biological mechanism(s) could parsimoniously account for chained rearrangements within a single cell cycle. The interpretation that chromoplexy coordinately generates chained rearrangements awaits experimental validation, which could involve FISH or chromosome conformation capture (3C) before and after inducing a predicted co-localizing event (e.g., testosterone exposure in prostate epithelial cells).

A unifying feature of chromoplexy-associated alterations is that they occur in a non-independent fashion; however multiple mechanisms may account for chromoplexy. Along these lines, our analyses have revealed distinctive patterns of chromoplexy in ETS^-, $CHD1^{del}$ tumors. Tumors with deletion of *CHD1* demonstrated an excess of intrachromosomal chained rearrangements and gene deletions, with DNA breakpoints enriched in GC-poor, late-replicating and non-expressed DNA. These tumors showed abundant, clustered rearrangements often affecting only one or two chromosomes with two alternating copy number states, perhaps indicating a chromothripsis-like process.

In contrast, chromoplexy in ETS^+ tumors differed in the aggregate from chromothripsis in several critical ways. For example, single events joined DNA from dispersed regions of six or more chromosomes in multiple tumors, whereas chromothripsis frequently involves focal rearrangement of one or two chromosomes (Forment et al., 2012). Overall, chromoplexy appears more prevalent in ETS^+ prostate cancer than chromothripsis is in any neoplasm (Forment et al., 2012; Stephens et al., 2011). Chromoplexy frequently involves fewer rearrangements than the “catastrophic” chromothripsis defined by Stephens et al., but may continue throughout tumor development. Our analysis of breakpoint locations in ETS^+ tumors suggests that chromoplexy in this setting may be linked to proposed transcriptional DNA-damaging processes (Lin et al., 2009), potentially related to androgen receptor signaling. Our findings align with the observation that ERG-overexpressing cancer cells accumulate DNA damage and are sensitive to poly ADP-ribose polymerase inhibition (Brenner et al., 2011). Chromoplexy is often active

prior to ETS gene fusions, however, and gave rise to *ERG* fusions in many instances. Ongoing analyses of cancer genomes may further elucidate mechanisms of chromoplexy and determine whether *CHD1* loss might also denote PARP-dependency, given the high degree of intrachromosomal rearrangement in this context.

Whole genome analysis also clarified the chronology of oncogenic events in prostate cancer progression, driven in part by chromoplexy. Genome-wide sequence coverage of germline SNPs allowed us to identify DNA lesions that arose after the founder clone was established. Subsequently, we demonstrated a progression of events within primary tumors that expands upon array-based SCNA co-occurrence studies (Demichelis et al., 2009). A consensus path of tumor evolution begins with events such as loss of *NKX3-1* or fusion of *TMPRSS2* and *ERG*. The path proceeds with the loss of *CDKN1B*, *TP53,* and *PTEN,* among other progression-associated lesions. We found that the histological grade of cancer may partially reflect its progression down this path.

A continuum model for tumor evolution

Tumorigenesis is classically understood to progress by a gradual accumulation of oncogenic alterations in the genome of a pre-cancerous cell. This textbook view was recently challenged by the discovery of chromothripsis, in which catastrophic rearrangements are incurred by “shattering” and reassembly of focal regions of the genome (Forment et al., 2012; Rausch et al., 2012; Stephens et al., 2011).

We propose an expanded model for the evolution of prostate cancer, which may also apply to other cancers (Figure 3.7). As classically understood, passenger and driver alterations can accumulate in a cancer genome gradually over numerous cell divisions, via point mutations, simple translocations and focal copy-number alterations. On the opposite end of the spectrum, extreme instances of chromothripsis can induce massive (albeit relatively localized) DNA damage at once, often with oncogenic consequences (Rausch et al., 2012; Stephens et al., 2011). Between these two extremes lies a broad continuum across which chromoplexy may often restructure cancer genomes. We propose that oncogenic events along this continuum reflect “punctuated” tumor evolution, drawing an analogy from the observation that

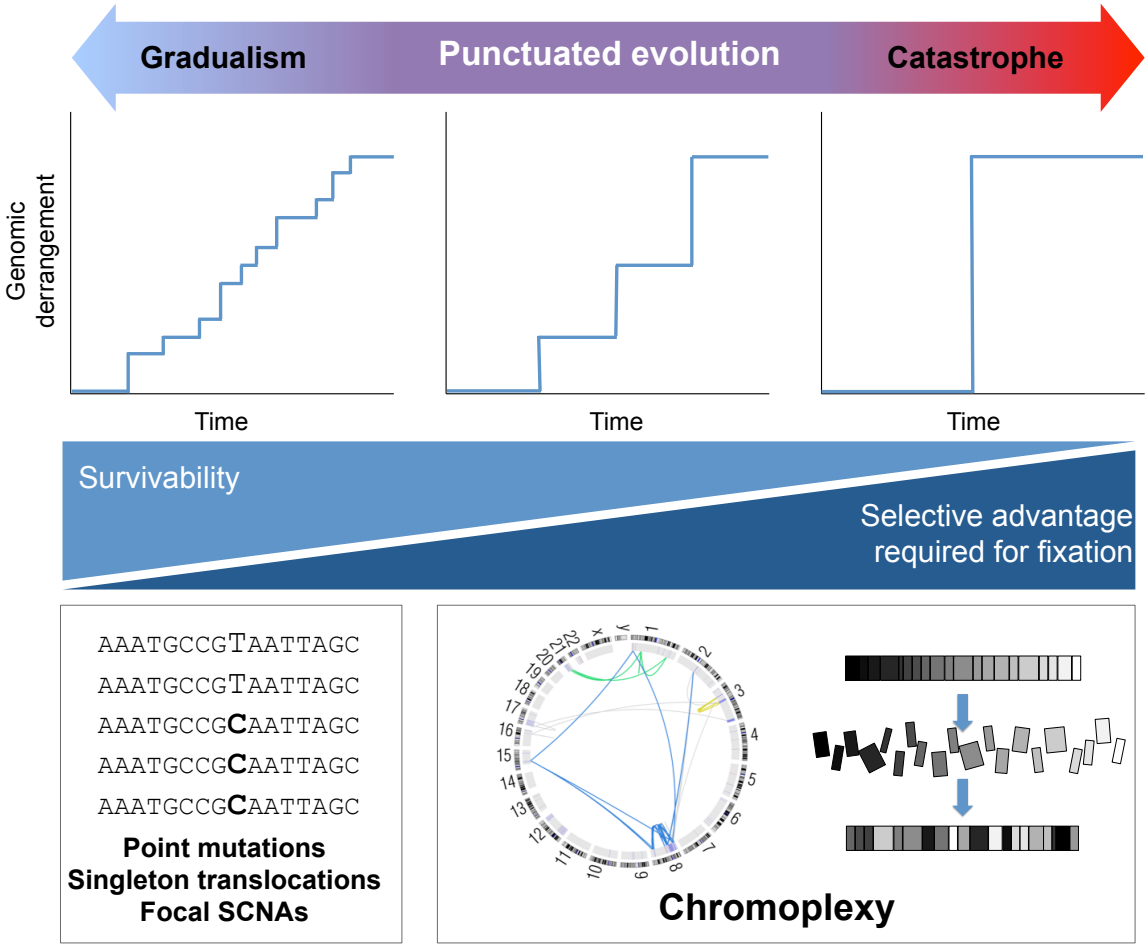


Figure 3.7. A continuum model for the genomic evolution of prostate cancer

Oncogenic aberrations may accumulate in cancer genomes gradually (left), by punctuated progression (middle) or in a single catastrophic event (right). Chromoplexy-associated rearrangements and deletions induce a modest to large degree of genomic derangement over several successive events. As indicated at bottom, larger-scale rearrangements that affect broader swaths of the genome may be more difficult for a cell to survive, and may tend to require co-occurring oncogenic lesions to become fixed in a tumor.

punctuated evolution of species may occur rapidly between periods of relative mutational equilibrium (Gould, 1977). By analogy, a tumor genome may sustain considerable damage over several sequential and punctuated events. Importantly, this framework accords with the observation that chromoplexy events (1) are common, (2) may involve a wide-ranging number of rearrangements, and (3) may continue after cancer-initiating lesions such as *NKX3-1* deletion (Figure S3.7).

A cancer might operate at any point along the continuum of progression at a given time. Tumors that develop primarily at the "catastrophic" end may require fewer events and could progress more quickly, because each such event could disrupt multiple cancer-constraining processes. At the same time, catastrophic events that cover diffuse genomic territory are more liable to disrupt essential or beneficial genes, thus imparting a selective disadvantage to (pre)malignant clones that sustain such events. Consequently, the model predicts that survivable chromoplexy (particularly near the catastrophic regime) is likely to involve oncogenic alterations that compensate for the incidental inactivation of essential genes (Figure 3.7). This prediction accords with the observation that most tumors show disruption of one or more putative prostate cancer genes within a chain. Moreover, this model raises the possibility that disruption of putative cancer genes by chromoplexy may heighten the probability that such genes represent "driver" events for that particular tumor. If so, this framework may hold important implications for the use of whole-genome sequencing in diagnostic and clinical studies.

In summary, this study highlights the potential for WGS data to capture aspects of the "molecular archeology" of cancer development that are missed by gene- or exome-level sequencing. The characterization of clonal progression and chromoplexy in emerging large panels of cancer genomes may provide insights about cancer initiation and progression, with implications for cancer detection, prevention and therapy.

Acknowledgements

DNA sequencing was performed by the Broad Institute Genome Sequencing Platform. The clonality of genomic lesions was assessed by Davide Prandi and Francesca Demichelis. Alessandro Romanel

analyzed the PPI network connectivity of rearranged genes. Yotam Drier and Olivier Elemento performed the breakpoint enrichment analyses.

Methods

Description of the tumor cohort

Prostate cancers analyzed in this study originated from two of the cohorts described in Chapter 3 (Barbieri et al., 2012) (Weill Cornell Medical College (WCMC; New York, NY) and Uropath Pty Ltd. (Perth, Australia), a provider of banked urological tissues). All prostate cancer samples were collected under an Institutional Review Board-approved protocol with the informed consent of patient donors. Sixteen tumors were characterized by exome-sequencing in a previous study (Table S3.1) (Barbieri et al., 2012). Previous analyses of SNP data from these cohorts indicated that patients were primarily of Caucasian ancestry (Barbieri et al., 2012). Primary adenocarcinomas were removed prior to any additional treatment for prostate cancer, including radiation therapy, brachytherapy or hormone ablation therapy. The two NEPC samples were reviewed by the study pathologists and confirmed as neuroendocrine carcinomas of prostatic origin based on clinical history and/or presence of ERG fusion (PR-7520). Immuno-histochemistry was negative for PSA and positive for the neuroendocrine marker synaptophysin in both cases.

Chromosomal copy number profiling

Segmented copy number profiles were generated from Affymetrix SNP 6.0 human microarray data as described in Chapter 3 (Barbieri et al., 2012). Sites of significant recurrent copy number alterations were identified by GISTICv2 (Beroukhim et al., 2010), with a $\log_2$ threshold of +/- 0.1 for amplification/deletion signals.

Sequencing data generation

WGS library construction

Libraries were prepared as described previously (Fisher et al., 2011) with slight modifications. First, the genomic DNA input into shearing was reduced from 3µg to 100ng in 50µL of solution. In addition, for adapter ligation, Illumina paired-end adapters were replaced with palindromic forked adapters with unique 8 base index sequences embedded within the adapter. Size selection was then performed using Sage Bioscience's Pippin Prep, with a target insert size of either 340bp or 370bp +/- 10%.

Following sample preparation, libraries were quantified using quantitative PCR (KAPA Biosystems) with probes specific to the ends of the adapters. This assay was automated using Agilent's Bravo liquid handling platform. Based on qPCR quantification, libraries were normalized to 2nM and then denatured with 0.1 N NaOH using Perkin-Elmer's MiniJanus liquid handling platform.

RNA-Seq library construction

RNA was isolated using a Dynabeads® mRNA Purifcation Kit (Life Technologies). Two rounds of poly-A selection (with bead regeneration) were performed to achieve rRNA contamination of less than 10%, as assessed by the Bioanalyzer mRNA Pico program (Agilent). Eluate was treated with DNase (TURBO DNA-free™ kit, Ambion) at 37°C for 30 minutes then immediately cleaned using RNAClean XP beads (Agencourt). RNA was fragmented in Fragmentation Buffer (Affymetrix) at 80˚C for 4 minutes. First- and second-strand cDNA synthesis were performed with SuperScript Double-Stranded cDNA Synthesis Kit (Life Technology). Library construction proceeded as described previously (Fisher et al., 2011), except that SPRI beads were used in the end-repair cleanup and standard paired-end adapters were replaced with barcoded adapters each containing a unique 8-base index sequence. After adapter ligation, two sequential cleanups were performed to remove adapter dimers, followed by 8 cycles of cDNA PCR amplification and SPRI cleanup. Before sequencing, samples were pooled and normalized according to qPCR results.

Cluster amplification and sequencing

Cluster amplification of denatured templates was performed according to the manufacturer's protocol (Illumina) using HiSeq v3 cluster chemistry and flowcells. Flowcells were sequenced with 101-bp paired end reads on a HiSeq 2000 using HiSeq v3 Sequencing-by-Synthesis Kits and analyzed using RTA v.1.12.4.2.

Genome sequence analysis

Sequencing data management and processing

A BAM file was generated for each sample from Illumina sequence reads using the Picard pipeline (http://picard.sourceforge.net/). Reads were mapped to the NCBI Human Reference Genome GRCh37 (hg19) with the Burrows-Wheeler Aligner (BWA) (http://bio-bwa.sourceforge.net). (BAM files from WGS data, as well as RNA-Seq and SNP array data were deposited in the database of Genotypes and Phenotypes (dbGaP; phs000447.v1.p1))

The cancer genome analysis pipeline known as Firehose (Principal author D. Voet) was used to manage and coordinate analysis of WGS data. Firehose submits input files and parameters to GenePattern (DePristo et al., 2011), which executes a series of analyses to verify data quality and detect somatic alterations by comparing tumor and normal sequences.

Quality control

We employed several quality control modules to monitor for contamination or potential sample mix-ups. To ensure that tumor- and normal-DNA were properly matched for a given individual and free of contaminating human sequences, we generated SNP fingerprints from 24 highly polymorphic sites for each sequencing lane. Lanes with outlier fingerprint genotypes for a given individual were discarded. In

addition, we used the ContEst algorithm (Cibulskis et al., 2011) to analyze homozygous non-reference SNPs to estimate levels of contamination with foreign human DNA, and required that samples demonstrate >95% concordance.

Normal DNA sequences were assessed for admixture with cancer DNA by examining copy number profile estimates based on sequence coverage in 100kb bins across the genome. Three normal tissue samples (PR-07-3258, PR-09-3983 and P05-2709) demonstrated low-level contamination with tumor DNA, based on similar patterns of DNA gains and losses between tumor and normal in a pair. In these cases, the histologically benign prostate tissue used as a source of normal DNA likely contained neoplastic or pre-neoplastic cells. The detection of somatic alterations in these samples was therefore limited. We specifically analyzed discarded rearrangement calls from these samples for prostate cancer-associated fusions, and identified cases in which the *TMPRSS2-ERG* fusion was filtered out due to its presence in normal (see below).

Detection of chromosomal rearrangements

Detection of somatic rearrangements was performed using the dRanger algorithm (Berger et al., 2011) to identify sequence reads from paired ends that map to the reference genome with unexpected orientations or intervals between read pairs. Candidate rearrangements were identified from clusters of such reads. They were then assigned a score by multiplying the number of chimeric reads supporting the fusion by a quality multiplier between 0 and 1. The quality multiplier takes into account the following factors: (1) the fraction of nearby reads with a mapping quality of zero; (2) the number and diversity of other discordant pairs in the vicinity of the breakpoints; and (3) the standard deviation of the starting positions of the supporting read pairs. Rearrangements with score of 4 or greater that were absent from the corresponding normal and from an extended panel of 176 non-cancerous genome sequences were classified as high-confidence. Rearrangements were categorized as deletions, inversions, interchromosomal translocations or tandem duplications based on the locations and strand directions of reads at fusion breakpoints.

In three cases, the *TMPRSS2-ERG* fusion was detected but filtered out, either due to low levels of tumor contamination in the adjacent prostate tissue used as a normal comparator (PR-09-3983, P05-2709), or to an abundance of breakpoints at the locus that resulted in a low quality multiplier (PR-STID0000000415). The *TMRPSS2-ERG* fusion was confirmed by fluorescence in situ hybridization (FISH) in these and all other fusion-positive cases.

Some loci that were rearranged in the context of a chain harbored many breakpoints that decreased the rearrangement quality score and caused the rearrangement to be rejected, despite support from multiple tumor reads and the absence of the rearrangement in normal DNA. To improve our ability to detect chains in these situations, we adjusted the parameters of dRanger so that rearrangements were considered if they demonstrate five supporting reads in tumor DNA, no reads in the panel of normal genomes and a score of 1 or greater. Rearrangements falling into this category were retained in the final dataset only if they were assigned to a chain.

Breakpoint fusion junctions were mapped to base-pair resolution where possible using the BreakPointer algorithm (Drier et al., 2012). BreakPointer searches for read pairs where one read maps near a breakpoint and the pair mate partially overlaps with the fusion junction, or fails to align anywhere. These unmapped reads are subjected to a modified Smith-Waterman alignment procedure with the ability to jump between the two reference sequences at the most fitting point. BreakPointer mapped the breakpoints to base pair resolution in 94% of the 5596 high-confidence rearrangements. In these cases, sequence homology at fusion junctions and any foreign sequence insertions were annotated.

Rearrangements were annotated with transcript information from the UCSC Genome Browser's UCSC Genes track (Table S3.3C) (Fujita et al., 2011) and illustrated using Circos (http://mkweb.bcgsc.ca/circos)

Identification and annotation of point mutations

We used the MuTect and IndelLocator algorithms to identify point mutations and small insertions/deletions (indels), respectively, as described in Chapter 3 (Barbieri et al., 2012). Point mutations and indels were annotated with information about relevant genes, transcripts, proteins and

other features using publicly available databases. A set of reference transcripts was compiled for annotation from the UCSC Genome Browser's UCSC Genes track as provided in the TCGA General Annotation Files (GAF) hg19 June 2011 bundle (https://tcga-data.nci.nih.gov/docs/GAF/). Variants were also annotated using the following resources: dbSNP build 134 (Sherry et al., 2001), UCSC Genome Browser's ORegAnno track (Griffith et al., 2008), UniProt release 2011_09 (Consortium, 2011) and COSMIC v55 (Forbes et al., 2011).

Validation of somatic mutations and rearrangements

Mutation validation from transcriptome sequences

We assessed 818 somatic point mutations covering annotated transcripts in RNA-Seq data from 20 tumors profiled by transcriptome sequencing (Table S3.3A). Of the mutated sites, 92 were covered by 40 or more RNA-seq reads and present in WGS reads at an allele fraction of 0.2 or greater. Of these mutations, 84 (91%) showed at least two reads supporting the alternate allele.

Validation of somatic rearrangements

Rearrangements were validated by two approaches. We assessed a set of 73 rearrangements, enriched for events affecting cancer genes, by PCR and deep sequencing on a MiSeq instrument (Table S3.3C). Reads from tumor and normal DNA were aligned to a custom genome that contained the hg19 reference genome along with sequences of all predicted somatic fusion junctions across samples. Rearrangements were classified as somatic if tumor, but not normal alignments, showed multiple high-quaity reads spanning the predicted fusion junction.

In addition, we selected 76 chromoplexy-associated rearrangements for validation by PCR alone (Figure S3.5, Table S3.3C). Primers were designed to amplify approximately 200bp containing the predicted fusion junction. Rearrangements were annotated as somatic if a band of the predicted size was amplified from tumor DNA but not from normal DNA.

Fluorescence in situ hybridization validation of rearrangements

ETS rearrangement was assessed using break-apart assays for *ERG* and *ETV1* as described previously (Berger et al., 2011). To assess genomic deletion, gene fusion and disruptive translocations, we used locus-specific dual-color FISH assays following a previously described approach (Berger et al., 2011; Perner et al., 2006). At least 50 nuclei were evaluated per tissue section using a fluorescence microscope (Olympus BX51; Olympus Optical). The following probes were used for FISH assays:

Locus	**BAC #**
CHD1	RP11-58M12
CHD1 Reference (5p13.1)	RP11-429D13
GSK3B 3'	RP11-59M4
GSK3B 5'	RP11-113H22
JAK1 3'	RP11-1061K17
JAK1 5'	RP11-76O023
JAK2 3'	RP11-274A3
JAK2 5'	RP11-259N10
CRKL 3'	RP11-76I4
CRKL 5'	RP11-1152E2
MAPK1 3'	RP11-317J15
MAPK1 5'	RP11-179H3

PTEN	CTD-2047N14
PTEN Reference (10q25)	RP11-431P18
FOXP1	RP11-410B2
FOXP1 Reference (3p11)	RP11-91M15
BRAF 3'	RP11-248P7
BRAF 5'	RP11-248O23

Protein-protein interaction (PPI) analysis of somatically rearranged genes

To identify gene rearrangements of potential biological consequence in Figure S3.1, we searched for recurrently rearranged genes whose protein products occupy central positions in interaction networks. To assess protein-protein interaction (PPI) network centrality, we considered the product of two measures of degree centrality and betweenness centrality:

1. **Degree centrality:** Given a protein p and a PPI network N, the index Degree(p,N) measures the number of interactions incident upon p. The index is normalized by dividing D(p,N) with the maximum index in the network.

2. **Betweenness centrality:** Given a protein p and a PPI network N, the index Betweenness(p,N) measures the number of shortest paths from all proteins to all others that pass through protein p. The index is normalized by dividing Betweenness(p,N) with the maximum index in the network.

We assessed centrality with the STRING database (Szklarczyk et al., 2011), and considered two other databases for independent support (Human Protein Reference Database (HPRD) (Prasad et al., 2009) and I2D (Brown and Jurisica, 2007)). The top quartile of centrality indexes in the entire network of 18,583 proteins was significantly enriched with protein products of the 397 genes with rearrangements in

more than one sample (p = 2 x 10^{-3}).

For rearranged genes that scored highly in the centrality analysis, we assessed gene expression levels in the subset of transcriptome-sequenced samples using RSEQtools (Habegger et al., 2011) (Table S3.4). To evaluate the effects of the rearrangements on gene transcription, we noted genes that were expressed in the bottom or top tenth percentiles in samples harboring rearrangement of the locus compared to all other tumors.

Detection of chained rearrangements and deletions

Overview of the ChainFinder algorithm

ChainFinder analyzes somatic DNA rearrangements from WGS data (e.g., deletions, inversions or translocations) and infers whether the rearrangement likely occurred in the context of a "chain" with two or more other rearrangements. Chained rearrangements are identified by searching for sets of breakpoints that are distributed about the genome in a configuration that would be improbable if the rearrangements had occurred independently of one another. The ability to detect chains is enhanced by also considering copy number profiles for signatures of chained rearrangements.

ChainFinder is implemented in MATLAB, and formulates the detection of rearrangement chains as a graph theory problem, in which breakpoints are treated as nodes that may be inter-connected by graph edges (Figures 3.3A and S3.3A). Edges connect pairs of breakpoints that are either (1) somatically fused to each other (2) involved in two distinct rearrangements that are unlikely to have arisen independently or (3) at either end of a deletion bridge. An initial graph is constructed by searching for sets of breakpoints and associated deletions for which the independent model can be rejected after correction for multiple hypothesis testing (see below). The initial graph is then refined by considering any alternative valid assignments of breakpoints and deletion segments to deletion bridges. In the final graph, breakpoints connected by edges correspond to collections of rearrangements that may have arisen concertedly in the context of a chain.

These steps are described in detail in the following sections and diagramed in Figure S3.3A.

Assessment of adjacent breakpoints

Each pair of breakpoints joined by a somatic DNA fusion is first connected by an edge on the graph (Figure S3.3A). For each pair of neighboring breakpoints on the reference genome within 1Mb of each other, the probability of two breakpoints arising independently within the observed distance of one another (P_{XY}) is calculated as follows. We assume that the probability of a DNA breakage event per nucleotide is uniform near the breakpoint and equal to μ_{local}. The probability of a second event *not* occurring within a distance *L* from the reference event (either upstream or downstream) is $(1-2\mu_{local})^L$. Therefore, the probability P_{XY} of observing a second breakpoint *Y* within distance L of an index breakpoint *X* is:

$$P_{XY} = 1- (1 - 2\mu_{local})^L$$

The rate μ_{local} is calculated based on (1) the number of breakpoints per base-pair observed in a given tumor (μ_{global}) and (2) the density of breaks near the rearranged locus across the panel of 57 prostate tumors (ρ):

$$\mu_{local} = \mu_{global}\,\rho$$

We estimate the breakpoint density ρ as a function of genomic location by dividing the genome into 1Mb windows and counting the number of tumors with one or more breaks within a given window (Figure S3.2B). Values of ρ are scaled uniformly such that the sum of μ_{local} across all windows is equal to μ_{global}.

For neighboring breakpoint pairs, P_{XY} is considered as a p-value for the hypothesis that the two breakpoints arose independently. Pairs of breakpoints are connected by an edge (assigned to the same chain) if the corresponding P_{XY} can be rejected with control of the false discovery rate at 10^{-2} (Benjamini, 1995).

Assignment of deletion bridges

Next, segmented copy number data are overlaid with breakpoint locations to identify rearrangement breakpoints that correspond to deletion events. This step connects breakpoints on the graph with edges corresponding to deletion bridges in cases where the breakpoints may have originated from the same DNA deletion event.

Each breakpoint is provisionally paired to a boundary of a deletion segment if the breakpoint lies within 8 SNP probes of the boundary (typically a span of several thousand base-pairs). Breakpoints at either boundary of a deletion segment are potentially joined by a deletion bridge if:

A. The breakpoints on either end of the deletion are not fused to each other; i.e., the deletion must correspond to a deletion bridge (involving two rearrangements) rather than a "simple deletion" (involving one rearrangement) (Figure 3.2A).
B. The sequencing reads supporting the breakpoints at either end of the deleted segment must "point towards" the deletion, such that the deleted sequence would lie directly downstream of the reads.

Edges are added to the graph to denote potential deletion bridges. In cases where pairs of breakpoints cannot be uniquely assigned to a single bridge, multiple interpretations are tested in a subsequent step (see "Finalization of the graph", below)

Evaluation of graph cycles

In some cases, P_{XY} is extremely small – for instance, when breakpoints from separate fusions map within several hundred base pairs of one another – and the breakpoints clearly did not originate independently. However, borderline cases often arise where P_{XY} is not sufficiently small to reject the independent model for two breakpoints unequivocally. In such cases, additional evidence that rearrangements were generated coordinately can be obtained by considering sets of breakpoints whose

nodes on the graph are contained within cycles (paths along edges that begin and end at the same node).

Each cycle is evaluated under the independent breakpoint model based on P_{XY} values for adjacent breakpoints within the cycle (Figure S3.3A). Specifically, all possible scenarios are considered by which one or more rearrangements within the cycle could have arisen independently. For example, three rearrangements involving six breakpoints in a hypothetical cycle (Figure S3.3B) could have arisen by the following (non-mutually exclusive) scenarios, where subscripted numbers in parentheses denoted rearrangements that occurred independently:

$$\{H_{(1)23}, H_{(2)13}, H_{(3)12}, H_{(1)(2)(3)}\}$$

This set of scenarios represents the independent model for the cycle, which encompasses all alternative possibilities to the breakpoints in the cycle arising coordinately (H_{123}).

ChainFinder considers the probability of detecting the independently generated breakpoints under each scenario within the observed distance of each other. Each scenario in the independent model requires that *two or more pairs* of adjacent breakpoints from separate rearrangements arise independently, in order to "split" the cycle into two or more separate events. Each such scenario can be expressed in terms of combinations of P_{XY} values from edges within the cycle corresponding to adjacent breakpoints (i.e., P_{ab}, P_{cd} and P_{ef}; Figure S3.3B)

$$\{(P_{ab}\,P_{cd}), (P_{ab}\,P_{ef}), (P_{cd}\,P_{ef}), (P_{ab}\,P_{cd}\,P_{ef})\}$$

As shown in Figure S3.3B, all scenarios involving three or more independent events require the co-occurrence of two or more scenarios involving only two events. Therefore, rejecting all scenarios involving only two events is sufficient to reject the independent model overall for the cycle. To assess all scenarios involving two independent events, ChainFinder tests the pairwise products of all P_{XY} values within a cycle (corresponding to all two-event scenarios) with control of the family-wise error rate (FWER; (Holm, 1979)) at 10^{-2} across all scenarios. Control of the FWER ensures that, if the independent model is rejected for a

cycle, there is a 1% chance that one or more of the independent rearrangement scenarios for the corresponding cycle were mistakenly rejected. All cycles for which the independent model is rejected are linked within a chain.

Finalization of the graph

Finally, the graph is refined by considering deletion bridges that could not be uniquely assigned. Although a single deletion bridge may exist that connects two breakpoints, frequently multiple interpretations are possible due to overlapping regions of deletion from separate alleles or distinct tumor subclones. In these cases, a single choice must be made from a set of mutually exclusive possible bridges. Bridges are mutually permissible only if the following conditions are met:

1. The bridges do not share the same breakpoints at either deletion segment boundary
2. If the bridges overlap, the deletion segment in the region of overlap must demonstrate a consistently lower copy number than segments outside the region of overlap.

ChainFinder tests permutations of mutually permissible bridges to find the combination that incorporates the most breakpoints into deletion bridges, because this solution best reconciles the copy number and rearrangement data. If a unique valid combination of bridges exists that maximizes the number of breakpoints in deletion bridges, the bridges are accepted and any distinct chains that they link are combined. If multiple optimal interpretations exist, only bridges that are included in all of these interpretations are kept.

The graph is finalized by removing any edge between neighboring breakpoints for which the independent-generation model could not be rejected. In addition, deletion bridge edges are retained in the graph only if the breakpoints on either end of the deletion arose non-independently (e.g., within a cycle).

Evaluation of genes disrupted in chains

Once chains have been assigned, a list of genes disrupted in each chain is compiled. Genes are included if they fall at least partially within a deletion bridge in the chain or within 10kb of a copy-neutral rearrangement in the chain. Circos plots are generated in which all rearrangements in a given chain are depicted in the same color (e.g., Figure 3.2B).

Assessment of false-positive rate with simulated tumor genomes

In order to test the false-positive rate of ChainFinder, we created "scrambled" tumor genomes by simulating the independent accumulation of rearrangements based on observed data. For each tumor, ten "scrambles" were created that combined rearrangements from other tumors. Each scramble contained the same number of rearrangements as the corresponding sequenced tumor. Any two rearrangements were combined in a scramble only if they were not part of the same chain from the same sequenced tumor. The scrambles served as "true negative" cases in which all rearrangements were generated independently, while preserving genome-position specific influences on breakage and fusion, since the data are drawn from observed rearrangements. Copy number profiles were simulated based on observed data as well. Segments of copy number alteration were generated that maintained (1) the number of breakpoints at the boundaries of potential deletion bridges and (2) the overall ratio of copy number gains to losses. The simulated rearrangement and copy number data were profiled with ChainFinder, and the proportion of breakpoints assigned to a chain was compared between observed and simulated data.

For each sequenced tumor, we also created ten simulations matched for rearrangement number and chromosomal connectivity. The rearrangement breakpoints were further matched to observed data with respect to (1) sequence coverage, (2) guanine and cytosine content of local sequence, (3) expression levels of nearby genes, (4) replication timing of DNA and (5) reference genome distance between breakpoints for intrachromosomal rearrangements (within 5%). Coverage was matched within 5x to the coverage near the observed breakpoint. Suitable locations for simulated breakpoints were identified by creating bins for the values of parameters (2) through (5) for each chromosome, and randomly choosing a location that falls within the same bin as the corresponding observed breakpoint. For each

feature (e.g., GC content), we created bins containing the bottom and top fifth percentiles across the chromosome. We then split the middle 90% evenly into three additional evenly spaced bins. Copy number profiles were simulated such that breakpoints at edges of deletion segments were preserved. In most cases where ChainFinder identified chains within simulated tumors, the simulations were too restrictive, so that the only matched location for a set of rearrangements in a chain was near to the location where they were observed.

Quantification of gene expression near rearrangement breakpoints

Expression was quantified in terms of gene-level FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values from 16 prostate tumor transcriptomes using CuffLinks (Trapnell et al., 2012). The transcription levels near rearrangements were estimated from median values of $\log_{10}$ (1+FPKM) across the tumor transcriptomes in 10kb windows on either side of the breakpoint. Where this window overlapped multiple genes, the largest FPKM value was used. For the analysis shown in Figure 3.4D, the statistical enrichment of chained breakpoints near highly expressed DNA in ETS-positive tumors was robust to exclusion of the *TMPRSS2* and *ERG* loci from expression level estimates.

Assessment of nuclear proximity of fused loci from Hi-C data

We sought to determine whether breakpoints involved in structural rearrangements are in close physical proximity in nuclei in which these breakpoints have not yet occurred. For this, we used filtered chromatin interaction data (Hi-C) from experiments performed in prostate epithelial cells (RWPE1) stably expressing a GFP reporter (RWPE1-GFP) (Rickman et al., 2012). To determine whether a set of breakpoint pairs are in close proximity, we defined a 1Mb window centered on each breakpoint and counted the Hi-C reads connecting the two windows for all breakpoint pairs. The average Hi-C read counts were determined separately for chained rearrangement breakpoints and for breakpoints that were not assigned to a chain for comparison. Rearrangements were further subdivided by ETS-status of the tumor in which they were observed.

We then compared the observed average Hi-C count to Hi-C counts that would be observed by chance if the breakpoint pairs were randomly distributed on the genome. We generated random sets of breakpoints matched to the observed breakpoints for intra-chromosomal distances, chromosomal distribution and short read mappability. We again defined 1Mb windows centered in the random breakpoints and counted Hi-C reads connecting each pair of simulated breakpoints. We repeated this analysis 1,000 times to generate a null distribution of average Hi-C read counts for random breakpoint pairs. To generate a p-value, we counted how many of the 1,000 sets of random breakpoints had an average Hi-C read count greater than or equal to the average read count for the observed breakpoints. Of note, only intra-chromosomal rearrangements were considered for this analysis, as inter-chromosomal breakpoints were supported by very few Hi-C reads even when considering large windows centered on the breakpoints.

Breakpoint enrichment analysis

Enrichment and depletion of breakpoints was assessed across the genome with respect to replication time, guanine/cytosine (GC) content and distance to transcribed genes. Observed distributions were compared to randomly generated distributions controlled for chromosome and coverage. First, nearby breakpoints (up to 2,500bp away) were consolidated into a single "event." For each event, 100,000 locations (one per iteration) were generated uniformly from all locations on the same chromosome having the same coverage. The genome was considered in the following bins: low GC (0-36%], medium GC (36%-45%] and high GC (45%-100%). Replication time was binned according to late/early ratio (Ryba et al., 2010) at (-∞,-0.8],(-0.8,0], (0,0.8],(0.8,∞). Changing the thresholds did not affect the essence of the results, other than losing sensitivity for very large or small bins (data not shown). For every bin we counted the number of breakpoints for both the observed breakpoints and the random breakpoints. All of these counts were used to compute nonparametric p-values (observed rates). Enrichment or depletion was determined by picking the lower of the one-sided p-values, and p-values were then corrected for multiple hypotheses by the Benjamini-Hochberg FDR procedure (Benjamini, 1995).

An analogous procedure was used to detect enrichment near ChIP-Seq peaks (Yu et al., 2010), except that bins were substituted with windows spanning 50kb on either side of each peak.

Quantification of tumor purity and subclonality

Prior to sequencing, estimates of tumor purity and ploidy were derived from Affymetrix SNP6.0 data using ABSOLUTE (Carter et al., 2012). These estimates were used to select high purity samples for whole genome sequencing (median purity 70%; ploidy range 1.84 – 2.21; Table S3.1A).

Analyses of tumor purity and subclonality from WGS data were performed by exploiting individuals' genotypes at polymorphic loci within somatically altered regions of the genome, using considerations related to previously described methods (Carter et al., 2012; Landau et al., 2013; Nik-Zainal et al., 2012). For each tumor sample included in the study, we estimated stromal DNA admixture and lesion clonality using CLONET (CLONality Estimate in Tumors; Prandi D. et al., manuscript in preparation). The approach behind CLONET and the MiSeq-based validation we performed are outlined hereafter.

For a tumor sample *TS* containing a mixture of N_{TS} normal (diploid) cells and T_{TS} tumor cells, the percentage of admixed normal cells is:

$$Adm(TS) = \frac{N_{TS}}{N_{TS} + T_{TS}}$$

Based on the above equation, the admixture can be estimated from sequencing reads covering a site of hemizygous deletion *s* as:

$$Adm_s(TS) = \frac{\beta_s(TS)}{2 - \beta_s(TS)}$$

where $\beta_s(TS)$ is the proportion of reads at locus *s* that originated from normal cells in TS.

In order to calculate *Adm*(*TS*), we first selected informative heterozygous SNPs within regions of somatic deletion that were identified from copy number array data. For each hemizygous deletion *H*, we considered the distribution of the allelic fractions (i.e., the fraction of reference sequence reads) from selected SNPs within *H*. Using "particle swarm optimization" (Kennedy, 1995) we calculated a composite value $\beta_H(TS)$ for the deleted region that best accounted for the observed distribution of allelic fractions at each heterozygous SNP *s* across the region. For every deletion, a value $Adm.apparent_H(TS)$ was computed that describes the apparent admixture at that locus. $Adm.apparent_H(TS)$ reflects both stromal admixture and potentially subclonal lesions that increase the apparent level of normal DNA at that locus. For each hemizygous somatic deletion *H,* the values of $Adm.apparent_H(TS)$ were grouped if the difference between the values could be explained by the estimation error determined by simulation-based error estimations. The smallest mean value of $Adm.apparent_H(TS)$ across a set of grouped deletions was taken as the candidate value of *Adm*(*TS*).

Estimates of cancer DNA purity by this procedure are listed in Table S3.1 and compared to estimates from the same tumors by ABSOLUTE run on SNP array data. The estimates were highly consistent across the samples ($R^2 = 0.99$; $p < 10^{-4}$) with the exception of two samples (PR-STID0000002682 and PR-07-360), where stromal admixture was detected in WGS data but not SNP array data.

We analyzed the clonality of gene deletions based on normalized $\log_2$ ratios of tumor and normal WGS sequence coverage, after correction for the estimated normal DNA admixture in tumor samples. Deletions where *Adm* and $Adm.apparent_H$ differ beyond the error estimation are potential sub-clonal lesions. We estimated the percentage of tumor cells that harbor a somatic hemizygous deletion *H*, i.e., the clonality of *H*, as:

$$Clonality_H(TS) = \frac{1 - Adm.apparent_H(TS)}{1 - Adm(TS)}$$

In the case of a 100% clonal hemizygous somatic deletion *H* the value of $Clonality_H(TS)$ is 1, as $Adm.apparent_H(TS)$ equals *Adm*(TS); otherwise $Clonality_H(TS)$ is less than 1. In the presence of high coverage, small variations in $Clonality_H(TS)$ can demonstrate differences in sub-clonality along a

continuous scale. Here, in order to avoid false positive calls for borderline subclonality, we adopted a conservative approach and only considered two classes of deletions: clonal ($Clonality_H(TS) \geq 0.8$) and subclonal ($Clonality_H(TS) < 0.8$). After *Adm* was calculated,we executed a similar procedure to estimate the clonal status of somatic homozygous deletions and point mutations.

The sensitivity of clonality detection depends upon the number of heterozygous SNPs within a deletion of interest and the depth of sequence coverage at these SNPs. We evaluated the uncertainty in clonality estimates as a function of these parameters by randomly sampling 1,800 simulations and averaging the difference between the true clonality and computed clonality for a given coverage and number of SNPs (Table S3.6). To ensure robust clonality calls, we considered only deletions with 20 or more informative SNPs with average sequence coverage of 20x (corresponding to a 5.4% estimation error). Table S3.7 lists the percentage of tumor cells found to harbor a specific lesion together with the associated uncertainty range.

Validation of clonality estimates

To assess our ability to estimate apparent DNA admixture from our WGS data, we generated independent validation data for a set of 18 aberrant genes with four heterozygous SNPs each from seven tumor samples by PCR and deep sequencing (>65,000x coverage). The deep coverage provided a precise estimate of the ratio of alleles at SNP sites. Figure S3.6A compares the local apparent DNA admixture for the 18 genes computed using WGS data to the estimates computed using deep sequencing data ($R^2 = 0.85$, $p = 3.55 \times 10^{-8}$). The contingency table inset in Figure S6A demonstrates agreement between WGS- and deep sequencing-based calls of clonality status (Cochran test, p-value = 1).

Additional statistical analyses

Quantitative comparisons of groups (e.g. numbers of rearrangements or SCNAs) were conducted with the rank-sum Mann-Whitney test, unless indicated otherwise. Box plots indicate median values and middle quartiles.

(Page intentionally left blank)

CHAPTER 4

Transcriptional effects of prostate cancer-associated *SPOP* mutations and *CHD1* inactivation

Introduction

The genomic analyses of prostate cancer described in the preceding chapters have identified several genomic alterations that occur primarily or exclusively in ETS fusion-negative (ETS$^-$) prostate tumors. In particular, mutation of the ubiquitin ligase subunit gene *SPOP* is strictly mutually exclusive with ETS gene fusions**.** Similarly, the chromatin modifying enzyme gene *CHD1* is deleted or rearranged primarily in ETS$^-$ tumors that also harbor *SPO*P mutations. These observations suggest that *SPOP* mutation and deletion of *CHD1* may contribute to prostate carcinogenesis in a distinct molecular subset of ETS$^-$ prostate cancers. In vitro studies have demonstrated that inactivation of *CHD1* and mutation of *SPOP* lead to invasive and morphological changes (Barbieri et al., 2012; Huang et al., 2011); however the oncogenic mechanisms engaged by these perturbations are unclear.

In order to study the consequences of *SPOP* mutation and *CHD1* deletion, we assessed the transcriptional profiles of prostate epithelial cells upon expression of mutant SPOP or knock-down of *CHD1*. We employed the LHMAR prostate epithelial cell line, which expresses large T antigen, hTERT, c-Myc and the androgen receptor. While LHMAR cells lack tumor-forming ability when injected subcutaneously into nude mice, they can be rendered oncogenic by overexpression of oncogenes such as H-Ras or the PI3-Kinase subunit p110α (Berger et al., 2004). LHMAR cells therefore represent a "partially transformed" cellular model for prostate oncogenesis that is useful for assessing the transforming effects of somatic DNA alterations observed in human tumors.

Given the postulated rolls in transcriptional regulation of SPOP and CHD1 (Gaspar-Maia et al., 2009; Kwon et al., 2006; Li et al., 2011), we hypothesized that *SPOP* mutation and *CHD1* inactivation could induce transcriptional changes in LHMAR cells that reflect the molecular pathways engaged by these events in prostate cancer. We therefore used transcriptome sequencing (RNA-Seq) to profile gene expression in LHMAR cells with knock-down of *CHD1* or expression of prostate cancer-associated *SPOP* mutant alleles. This chapter describes our analysis of signatures associated with these perturbations and discusses mechanisms by which these events may drive ETS$^-$ prostate cancer.

CHD1 inactivation activates cellular DNA damage responses

To assess the effects of *CHD1* inactivation in prostate epithelial cells, we suppressed *CHD1* expression with three independent short hairpin RNAs (shRNAs) targeting non-overlapping regions of the gene. Knock-down was verified by Western blot (Figure S4.1). We performed RNA-seq and used CuffDiff 2.0 (Trapnell et al., 2012) to assess expression changes in 8,581 genes upon *CHD1* knock-down in LHMAR cells compared to a non-targeting control hairpin. All three hairpins caused largely concordant changes in transcription. For example, the top 5th percentiles of up- and down-regulated genes, overlapped significantly across the hairpins (Figure 4.1, left, $p < 10^{-4}$ for each pair-wise comparison). Similarly, significant overlap was observed among genes that were up- or down-regulated at least two-fold by each hairpin as well (Figure 4.1, right).

We used gene set enrichment analysis (GSEA; (Subramanian et al., 2005)) to search for curated gene sets or oncogenic expression signatures that were significantly up- or down-regulated with *CHD1* suppression. Eighteen gene sets were significantly overexpressed upon knock-down of *CHD1* at a false discovery rate (FDR) of < 10^{-4} (Table 4.1). Strikingly, the majority of up-regulated gene sets (13/18) pertained to cellular DNA damage response and p53 signaling (Figure 4.2). Previous studies have suggested a role for *CHD1* in maintaining genomic stability (Huang et al., 2011; Liu et al., 2012) and prostate tumor genomes with deletion of *CHD1* demonstrate a significant excess of chromosomal rearrangements (Chapter 3; Baca et al., 2013). Therefore, inactivation of *CHD1* may induce DNA damage or prevent repair of DNA double strand breaks (DSBs), perhaps as a consequence of aberrant chromatin compaction or impaired access of DSB repair machinery to heterochromatin. Along these lines, inactivation in the mouse germline of *Chd2*, a closely related homolog, disrupts DSB repair and confers susceptibility to leukemias. Thus, *CHD1* loss might promote tumor growth by facilitating the inactivation of additional tumor suppressor genes. This hypothesis will need to be explored experimentally in further detail, but is in line with the observation that *CHD1*-deleted tumors are enrichmed with other recurrent DNA deletions, such as loss of *PTEN* and *CDKN1B* (Chapter 3, Figure S3.4).

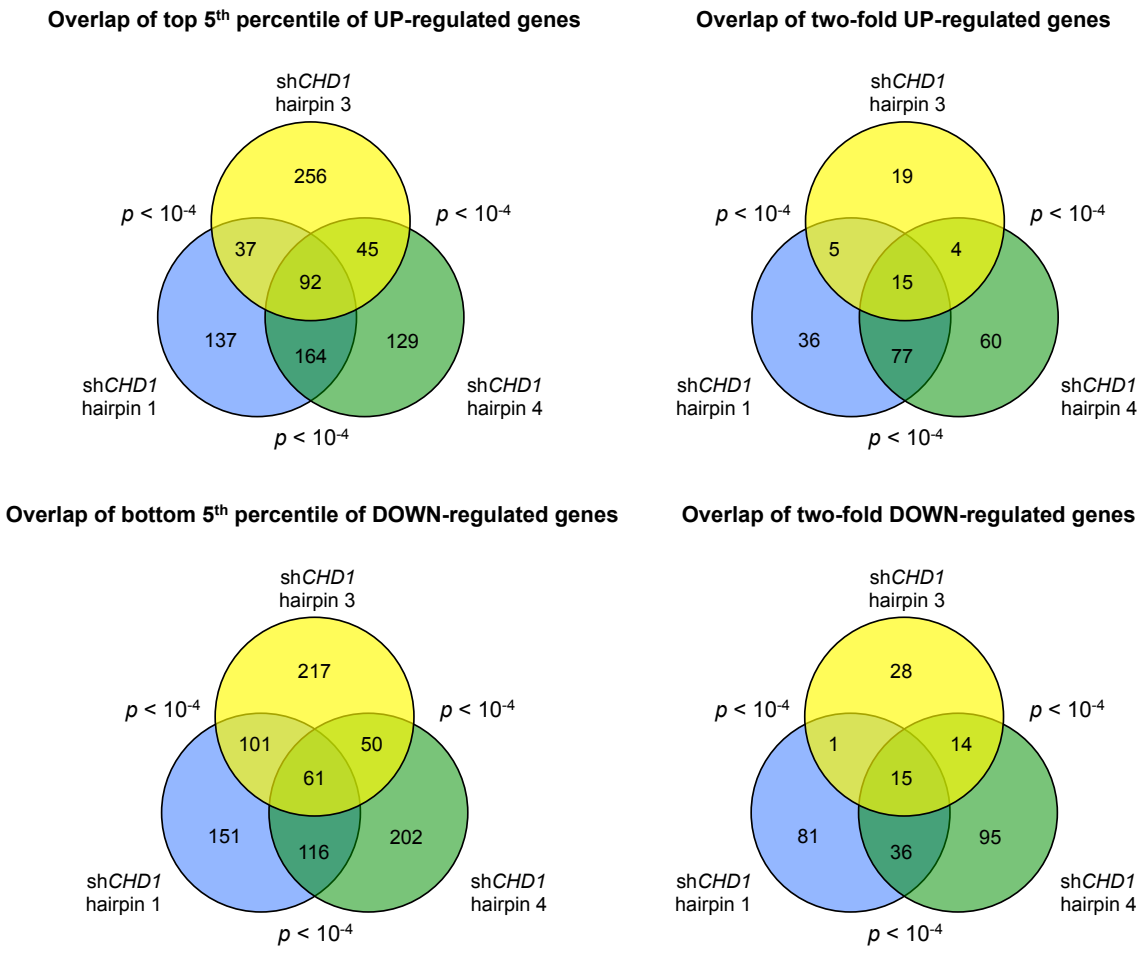


Figure 4.1. Knock-down of *CHD1* with three independent hairpins alters the expression of overlapping genes

The overlap of transcriptional responses upon *CHD1* knockdown by each of three hairpins was assessed from RNAseq data using CuffDiff 2.0. LHMAR cells with a non-targeting hairpin served as a comparator. Significant overlap was observed between the top and bottom fifth percentiles of up- and down-regulated genes with each hairpin. Similar overlap was evident in genes that were up- or down-regulated by twofold compared to the non-targeting hairpin control cells.

Table 4.1. Gene sets related to DNA damage and p53 signaling are up-regulated with knock-down of *CHD1*

Expression changes were averaged across LHMAR cell lines with *CHD1* knock-down by three independent hairpins compared to control shRNA. Gene set enrichment analysis was used to identify gene sets that are significantly up-regulated with *CHD1* knock-down. All gene sets with FWER p-values and FDR q-values < 10^{-4} are listed.

Gene set	Genes	Genes in enrichment signal	Normalized enrichment score	FDR q-value	FWER p-value
WARTERS_RESPONSE_TO_IR_SKIN	48	34	3.14	< 10^{-4}	< 10^{-4}
SMIRNOV_RESPONSE_TO_IR_6HR_UP	118	60	3.11	< 10^{-4}	< 10^{-4}
PID_P53DOWNSTREAMPATHWAY	97	46	2.98	< 10^{-4}	< 10^{-4}
KERLEY_RESPONSE_TO_CISPLATIN_UP	34	26	2.81	< 10^{-4}	< 10^{-4}
WARTERS_IR_RESPONSE_5GY	26	17	2.74	< 10^{-4}	< 10^{-4}
SCHAVOLT_TARGETS_OF_TP53_AND_TP63	14	13	2.56	< 10^{-4}	< 10^{-4}
DER_IFN_GAMMA_RESPONSE_UP	59	27	2.54	< 10^{-4}	< 10^{-4}
SANA_TNF_SIGNALING_UP	42	20	2.51	< 10^{-4}	< 10^{-4}
GHANDHI_DIRECT_IRRADIATION_UP	46	26	2.50	< 10^{-4}	< 10^{-4}
DER_IFN_ALPHA_RESPONSE_UP	59	23	2.47	< 10^{-4}	< 10^{-4}
GENTILE_UV_LOW_DOSE_UP	23	11	2.46	< 10^{-4}	< 10^{-4}
BRACHAT_RESPONSE_TO_CAMPTOTHECIN_UP	24	12	2.46	< 10^{-4}	< 10^{-4}
BRACHAT_RESPONSE_TO_METHOTREXATE_UP	19	10	2.44	< 10^{-4}	< 10^{-4}
KIM_GLIS2_TARGETS_UP	33	16	2.41	< 10^{-4}	< 10^{-4}
BRACHAT_RESPONSE_TO_CISPLATIN	18	9	2.41	< 10^{-4}	< 10^{-4}
PHONG_TNF_TARGETS_UP	39	22	2.40	< 10^{-4}	< 10^{-4}
ONGUSAHA_TP53_TARGETS	26	12	2.37	< 10^{-4}	< 10^{-4}
RPS14_DN.V1_UP	54	27	2.37	< 10^{-4}	< 10^{-4}

Exemplary gene sets up-regulated with *CHD1* knock-down

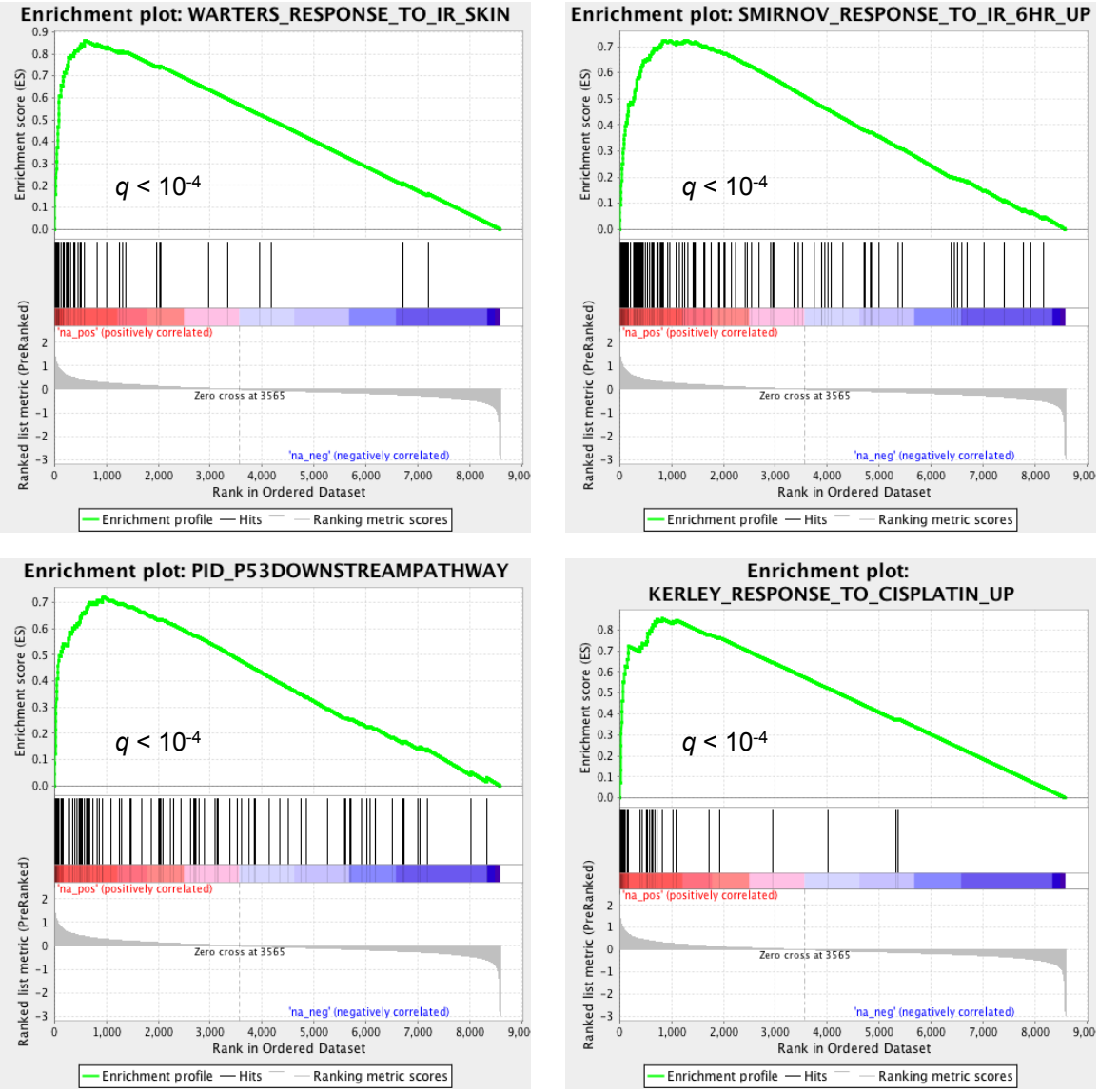


Figure 4.2. GSEA enrichment plots showing up-regulation of gene sets related to DNA-damage response and p53 signaling

FDR q-values are indicated for the enrichment of each gene set.

SPOP mutations activate transcriptional programs related to invasion, TGF-β signaling and polycomb repression

We also used the LHMAR prostate epithelial cell line to probe the effects of SPOP mutation on cellular transcriptional output. SPOP mediates the ubiquitination of several transcriptional regulators such as DAXX and SRC3 (Kwon et al., 2006; Li et al., 2011) and *SPOP* mutations often occur in the absence of oncogenic lesions in prostate cancer genes such as *TP53* or *PTEN* (Chapter 2). We therefore reasoned that *SPOP* mutation might activate oncogenic transcriptional programs. Furthermore, we hypothesized that distinct prostate cancer-associated *SPOP* mutations would exert similar effects, given their clustered distribution within the protein substrate binding pocket (Chapter 2). We therefore used RNA-Seq to assess whether two prostate cancer-associated SPOP mutants (F133L and Y87N) induce concordant changes in gene expression compared to wild-type SPOP.

To determine whether the two SPOP mutants had similar effects on transcription in prostate epithelial cells, we compared sets of genes with altered expression in both mutant-expressing cell lines according to our CuffDiff 2.0 analysis. We found significant overlap of genes up- or down-regulated by two-fold, or genes within the top fifth percentiles of differentially expressed genes (Figure 4.3). These results suggest that expression changes associated with SPOP mutation do not solely reflect “noisy” transcripts, and that the two mutants may have similar consequences at the transcriptional level.

Our differential expression analysis pointed toward several cellular processes and pathways on which mutant SPOP might impinge in prostate cancer. Using GSEA, we identified 152 gene sets that were significantly up-regulated with mutant SPOP at an FDR q-value < 10^{-4} (Table 4.2). Many of these gene sets could be grouped into shared cancer-related pathways or processes. For example, multiple gene sets were enriched pertaining to cellular invasion and epithelial to mesenchymal transition (EMT) (Acloque et al., 2009). Additionally, TGF-β signaling pathways (which mediate invasive growth in several cancers (Elliott and Blobe, 2005)) were up-regulated. These findings are consistent with the observation

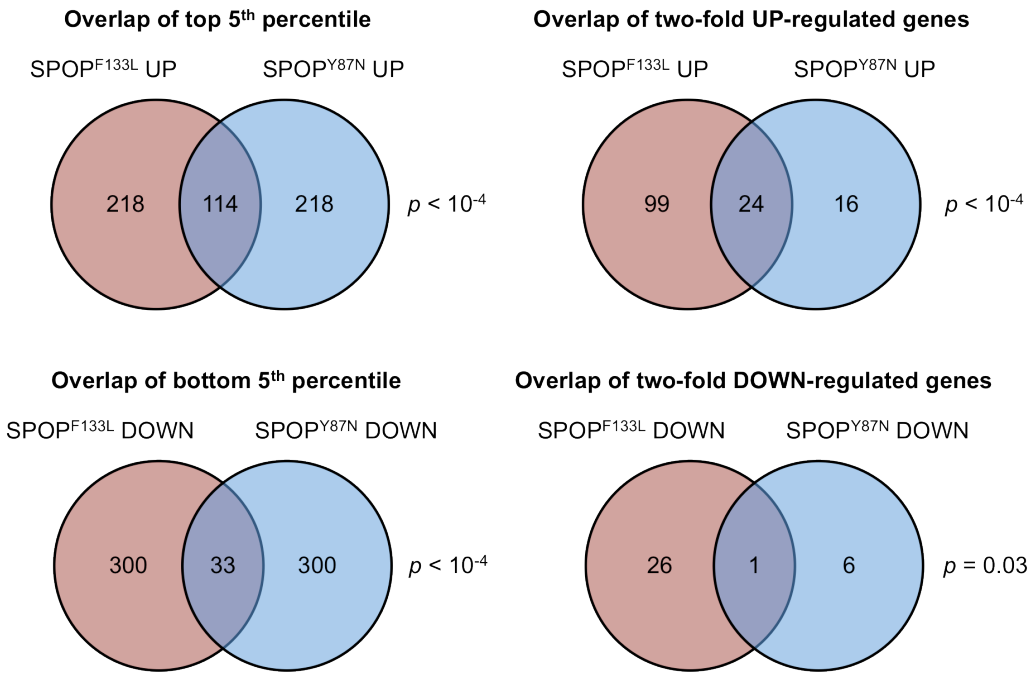


Figure 4.3. Two prostate cancer-associated SPOP mutations induce expression of an overlapping set of transcripts in prostate epithelial cells

Differentially regulated genes were identified by CuffDiff 2.0 analysis of RNAseq data from SPOP mutant (F133L and Y87N) and SPOP wildtype-expressing LHMAR cells. To assess the overlap of transcriptional responses between the two SPOP mutants, we identified genes that were up- or down-regulated by either mutant compared to wild-type SPOP.

Table 4.2. SPOP mutation induces transcriptional changes related to cell invasion, polycomb repression, TGF-β signaling and Wnt/β-catenin signaling

Enriched gene sets were identified by GSEA analysis of the average changes in gene expression with $SPOP^{F133L}$ and $SPOP^{Y87N}$ compared to wild-type SPOP. Selected gene sets with FDR q-values < 10^{-4} are depicted and grouped by common molecular pathways or processes.

Gene set	Genes	Genes in enrichment signal	Normalized enrichment score	FDR q-value	FWER p-value	Associated pathway(s)
GU_PDEF_TARGETS_UP	45	23	2.33	< 0.001	< 0.001	EMT/invasion
KEGG_CELL_ADHESION_MOLECULES_CAMS	35	26	2.26	< 0.001	< 0.001	EMT/invasion
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_UP	123	70	2.68	< 0.001	< 0.001	EMT/invasion
KEGG_ECM_RECEPTOR_INTERACTION	35	25	2.37	< 0.001	< 0.001	EMT/invasion
ROZANOV_MMP14_TARGETS_UP	129	46	2.30	< 0.001	< 0.001	EMT/invasion
WU_CELL_MIGRATION	90	48	2.29	< 0.001	< 0.001	EMT/invasion
ANASTASSIOU_CANCER_MESENCHYMAL_TRANSITION_SIGNATURE	25	20	2.27	< 0.001	< 0.001	EMT/invasion
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	29	24	2.17	< 0.001	0.002	EMT/invasion
CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_DN	299	116	2.16	< 0.001	0.003	EMT/invasion
JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	38	20	2.13	< 0.001	0.006	EMT/invasion
PID_INTEGRIN1_PATHWAY	34	25	2.33	< 0.001	< 0.001	Integrin
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	33	20	2.30	< 0.001	< 0.001	Integrin
PID_INTEGRIN3_PATHWAY	26	17	2.27	< 0.001	< 0.001	Integrin
MEL18_DN.V1_UP	64	46	2.79	< 0.001	< 0.001	Polycomb repressive complex
BMI1_DN_MEL18_DN.V1_UP	63	42	2.70	< 0.001	< 0.001	Polycomb repressive complex
BMI1_DN.V1_UP	69	47	2.67	< 0.001	< 0.001	Polycomb repressive complex
WIEDERSCHAIN_TARGETS_OF_BMI1_AND_PCGF2	35	24	2.36	< 0.001	< 0.001	Polycomb repressive complex
KONDO_EZH2_TARGETS	73	32	2.13	< 0.001	0.006	Polycomb repressive complex
NUYTTEN_EZH2_TARGETS_UP	485	204	2.13	< 0.001	0.007	Polycomb repressive complex
PLASARI_TGFB1_TARGETS_10HR_UP	75	35	2.38	< 0.001	< 0.001	TGFβ
VERRECCHIA_DELAYED_RESPONSE_TO_TGFB1	26	16	2.31	< 0.001	< 0.001	TGFβ
VERRECCHIA_EARLY_RESPONSE_TO_TGFB1	44	26	2.29	< 0.001	< 0.001	TGFβ
LABBE_TGFB1_TARGETS_UP	46	24	2.22	< 0.001	< 0.001	TGFβ
TGFB_UP.V1_UP	64	23	2.14	< 0.001	0.005	TGFβ
SANA_TNF_SIGNALING_UP	37	23	2.40	< 0.001	< 0.001	TNF
PHONG_TNF_RESPONSE_NOT_VIA_P38	191	87	2.35	< 0.001	< 0.001	TNF
ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND_TNF_UP	82	42	2.26	< 0.001	< 0.001	TNF
PHONG_TNF_RESPONSE_VIA_P38_PARTIAL	85	49	2.19	< 0.001	0.001	TNF
PHONG_TNF_TARGETS_UP	35	25	2.19	< 0.001	0.001	TNF
ONDER_CDH1_SIGNALING_VIA_CTNNB1	39	23	2.44	< 0.001	< 0.001	Wnt/β-catenin
ONDER_CDH1_TARGETS_2_UP	119	50	2.34	< 0.001	< 0.001	Wnt/β-catenin
LABBE_TARGETS_OF_TGFB1_AND_WNT3A_UP	51	26	2.22	< 0.001	< 0.001	Wnt/β-catenin; TGFβ

that expression of mutant SPOP, but not wild-type SPOP, increases the invasion of prostate cancer cell lines (Chapter 2).

Several pathways related to known or putative SPOP substrates were enriched as well. For example, polycomb repressive complex (PRC) signature genes were overexpressed in the setting of mutant SPOP, including targets of the BMI1 polycomb repressive protein (Table 4.2). This finding may be noteworthy given that BMI1 is a substrate of the SPOP-CUL3 ubiquitin ligase complex, and BMI1 is a key regulator of prostate stem cell self-renewal (Lukacs et al., 2010). Additionally, we recently identified the zinc finger protein WIZ as a potential SPOP substrate relevant to prostate cancer using a mass spectrometry-based screening approach. In LHMAR cells, F133L and Y87N mutant SPOP alleles significantly inhibit the ubiquitination of this protein, while wild-type SPOP increases ubiquitinated WIZ levels (J.P. Theurillat et al., manuscript in preparation). WIZ participates in the euchromatic histone H3K9 methyl-transferase complex with C-terminal binding protein (CtBP) (Ueda et al., 2006), which may regulate transcription of genes involved in Wnt/β-catenin signaling, EMT and stem cell maintenance (Di et al., 2013; Fang et al., 2006). Both CtBP and Wnt/β-catenin related pathways were up-regulated with expression of mutant SPOP. Furthermore, analysis of human tumor RNA-Seq data from 16 cases (two *SPOP* mutant and 14 *SPOP* wildtype) showed activation of CtBP target genes in *SPOP* mutant tumors. These observations raise the possibility that SPOP mutations might perturb CtBP signaling in human tumors via effects on ubiquitination of WIZ.

Discussion and conclusion

We have examined the transcriptional output of prostate epithelial cells upon inactivation of *CHD1* or expression of mutant *SPOP* to study how these perturbations may mediate oncogenesis in human prostate cancer. We observed transcriptional profiles suggestive of DNA damage and p53 signaling upon knock-down of *CHD1*, which may suggest that *CHD1* loss facilitates the accumulation of abundant SCNAs and rearrangements seen in *CHD1*-deleted tumors. Likewise, expression of mutant SPOP

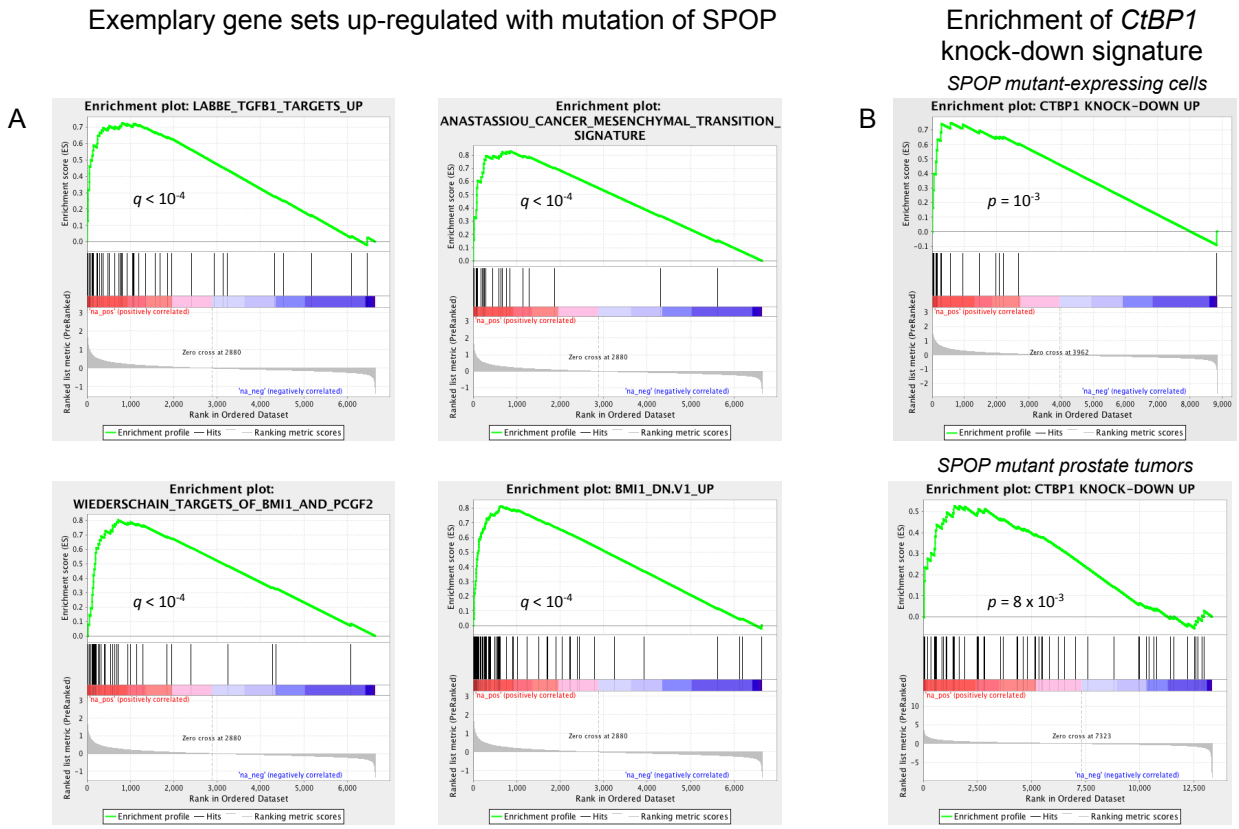


Figure 4.4. Enrichment plots of gene sets up-regulated with mutation of SPOP

(**A**) Enrichment plots of selected gene sets up-regulated in LHMAR cell lines with SPOP mutant expression (see Table 4.2).

(**B**) Both LHMAR cells (top) and human prostate tumor tissues (bottom) expressing mutant SPOP were assessed for up-regulation of genes that showed increased expression with knock-down of CtBP-1 in prostate cancer cell lines in a previous study (Wang et al., 2012).

appeared to activate specific transcriptional modules compared to wild-type SPOP. Analysis of enriched gene sets in prostate epithelial cells expressing mutant SPOP suggested that *SPOP* mutation may drive EMT- and invasion-related gene transcription, and may activate genes that are normally repressed by the polycomb repressive complex and CtBP complex.

Overall, these results suggest that *SPOP* mutation and *CHD1* loss dysregulate distinct molecular processes, and provide hypotheses for ongoing experimental studies. It should be emphasized that our transcriptome analyses provide only preliminary insight into oncogenic mechanisms, and are limited by several factors. First, it is not known whether the LHMAR cells reflect the pathophysiological cell of origin of human *SPOP*-mutant or *CHD1*-deleted prostate cancers. In addition, collaboration of other oncogenic alterations may be required to fully unmask the effect of cancer-associated *SPOP* mutations. Furthermore, it is not yet known whether the transcriptional changes observed with these perturbations are responsible for the outgrowth of *SPOP*-mutant or *CHD1*-deleted cancer cell clones in human tumors. Nonetheless, the results of these experiments highlight several avenues for experimental study. Ongoing experiments will leverage transgenic mouse models to corroborate our findings, and will assess larger sets of *SPOP*-mutant or *CHD1*-deleted tumors for the transcriptional changes observed in our LHMAR model.

Acknowledgements

RNA-Seq was performed by the Broad Institute Genome Sequencing Platform. Cell culture and western blot analysis of SPOP and CHD1 expression were performed by Jean-Phillippe Theurillat.

Methods

Cell culture

LHMAR cells were previously derived from primary prostate epithelial cells via stable expression of Large T antigen, *hTERT*, c-Myc and the androgen receptor (Berger et al., 2004). Three shRNA contstructs in the pLKO.1 vector targeting *CHD1* were virally transduced into LHMAR cells. Two days after infection, cells were selected in puromycin until no viable cells remained in a non-infected control plate (two days). Cells were then cultured for 7 days in RPMI 1640 and 10 million cells per condition were snap frozen in liquid nitrogen. The non-silencing hairpin was ordered from Sigma (SHC002 Sigma MISSION pLKO.1-puro Non-Mammalian shRNA Control Plasmid DNA). *CHD1*-directed hairpins were obtained from The RNAi Constortium (http://www.broadinstitute.org/rnai/public/) and were designed to target the following *CHD1* sequences:

Hairpin 1: GCGGTTTATCAAGAGCTATAA
Hairpin 3: GCGCAGTAGAAGTAGGAGATA
Hairpin 4: GCAGTTGTGATGAAACAGAAT

CHD1 knock-down was verified by western blot using goat polyclonal antibody (Novus Biological, NBP1-00168).

Wild-type and mutant SPOP constructs with a Kozac ACC-sequence were coned into the pCW107 plasmid with the mOrange fluorescence marker using Neh1 and Mlu1 restriction enzymes. 100% infection efficiency was confirmed by fluorescence microscopy. SPOP expression was confirmed by western blot using rabbit polyclonal antibody (Abcam, ab81163). Cells were cultured for a week in RPMI 1640 and 10 million cells were snap frozen in liquid nitrogen.

RNA-sequencing data generation

RNA-Seq, including library synthesis, sequencing and data processing were performed as described in Chapter 4 (Methods).

Analysis of differentially regulated genes and gene sets

We used CuffDiff 2.0 (Trapnell et al., 2012) to compare gene-level expression between cell lines. Transcript and gene annotations were obtained from Gencode Version 7 (http://www.gencodegenes.org). A "mask" file was used to exclude several categories of small RNA transcripts from consideration by CuffDiff 2.0, including transcripts annotated with "Mt_rRNA", "Mt_tRNA", "misc_RNA", "rRNA", "snRNA" or "snoRNA". Additionally, genes annotated as "PUTATIVE" or "NOVEL" were excluded. In general, CuffDiff 2.0 was configured to require a minimum of 200 reads from a gene to assess differential expression between conditions. For the analysis of CtBP1 knock-down signature enrichment, we decreased the minimum read requirement to 100 reads in order to increase the number of genes within the CtBP1 knock-down gene set that could be assessed by CuffDiff 2.0.

Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was used to assess enrichment of up- or down-regulated transcripts between cell lines. GSEA was applied to the list of transcripts assessed by CuffDiff 2.0 that was "pre-ranked" based on average change in expression between the compared conditions (e.g., *CHD1* hairpin versus control hairpin or *SPOP* mutant versus wild-type). Our GSEA analysis included all gene sets within MSigDB containing between 15 and 500 members under the categories "C2: curated gene sets" and "C6: oncogenic gene sets" (3,624 sets total). In addition, we constructed a custom gene set corresponding to potential CtBP1 targets from a list of genes that were up-regulated > 2-fold with knock-down of CtBP1 in DU145, PC3 and LNCaP prostate cancer cell lines in a previous study (Wang et al., 2012).

Statistical analysis

Two-sided p-values from Fisher's exact tests are indicated for comparison of up- and down-regulated transcripts between cell lines.

CHAPTER 5

Conclusions and future directions

The studies described in this thesis provide an early sketch of the landscape of somatic alterations in prostate cancer genomes. This research builds upon studies over the last two decades that have probed prostate tumor DNA at ever-finer resolution to uncover alterations that drive disease initiation, progression, and therapeutic resistance (Shen and Abate-Shen, 2010). By sequencing the protein-coding genes in a large panel of tumors, we were able to identify positive selection for mutations in genes that were not previously implicated in cancer, such as *FOXA1*, *MED12* and *SPOP*. By analyzing entire genomes from a smaller tumor panel, we gained insight into the clonal evolution of prostate cancers and discovered that genomic rearrangements may accumulate in a highly inter-dependent manner.

Novel putative prostate cancer genes implicated by tumor DNA sequencing

Our analysis of 112 prostate cancer exomes revealed novel significantly mutated genes that may point to important molecular processes in the pathogenesis of this disease. For example, recurrent mutations in the substrate-binding cleft of the SPOP ubiquitin ligase subunit may prevent turnover of oncogenic proteins by ubiquitin-mediated proteolysis. Along these lines, several instances of altered proteolytic homeostasis have been documented in prostate cancer. For example, overexpression of the Skp2 ubiquitin ligase promotes degradation of tumor-constraining proteins such as $p27^{Kip1}$ (Carrano et al., 1999; Yang et al., 2002). Additionally, the truncated form of ETV1 that is expressed in the context of the *TMPRSS2-ETV1* fusion escapes COP1-mediated ubiquitination and degradation, and may thereby accumulate to pathological levels (Vitari et al., 2011). Similarly, the accumulation of certain oncogenic proteins might drive *SPOP*-mutant prostate cancer, if the documented SPOP mutations prevent substrate binding as predicted.

The relevant targets of SPOP in prostate oncogenesis are not yet known. Several substrates of the SPOP-CUL3 complex have been identified, including the transcriptional repressor DAXX (Kwon et al., 2006), the hedgehog pathway transcription factor Gli (Zhang et al., 2009) and the histone variant protein MACROH2A (Hernandez-Munoz et al., 2005). Experiments are underway to assess the proteome-wide

changes in protein ubiquitination induced by *SPOP* mutations, and may shed light on the drivers of *SPOP*-mutant cancer. Preliminary results from transcriptome-sequencing studies described in Chapter 4 point to processes associated with invasion and metastasis as well as polycomb and TGF-β signaling, all of which have been implicated in prostate tumorigenesis (Ding et al., 2011; Lukacs et al., 2010; Ru et al., 2012). As discussed in Chapter 2, a dominant-negative effect of SPOP mutations seems likely given the cluster of mutated residues in SPOP, the absence of missense mutations or bi-allelic inactivation, and previous knowledge that SPOP functions in a homo-dimeric complex (Zhuang et al., 2009). Therefore, elucidating relevant SPOP substrates may not only inform our understanding of prostate cancer biology, but may also point to therapeutic targets that are upregulated due to decreased proteolytic turnover.

The transcription factor gene *FOXA1* was another novel cancer gene implicated by this work. FOXA1 functions in prostatic development (Gao et al., 2005) and facilitates lineage-specific transcription by binding to DNA enhancer elements (Lupien et al., 2008). The recurrent mutations that cluster within DNA-interacting residues of FOXA1 may disrupt transcriptional regulation by this protein. *FOXA1* adds to a growing list of prostatic developmental genes that are somatically altered in prostate cancer. Two well-characterized examples include *NKX3-1*, a regulator of prostatic developmental transcription that is frequently deleted, and the androgen receptor, which coordinates androgen-mediated transcriptional programs and is mutated in castration-resistant disease. Collectively, these alterations suggest that prostate oncogenesis involves the halted natural development and differentiation of prostate epithelial cells – a concept that has been explored in other cancers (Chou et al., 2010; Ferrero et al., 1982; Yang and Weinberg, 2008). Interestingly, *FOXA1* mutations and *NKX3-1* deletions were strictly clonal in our tumor panel, suggesting that they arise within a common ancestor of sequenced tumor cells and occur early in disease. These alterations might revert incipient prostate cancer cells into a de-differentiated state that renders them susceptible to subsequent oncogenic insults such as *PTEN* or *CDKN1B* inactivation (Chapter 3).

An important finding of our studies is that many genes underwent recurrent and/or oncogenic alterations in only a small subset of tumors. For example, several genes such as *IDH1*, *HRAS*, *KDM6A* revealed cancer-associated mutations in < 3% of samples (Chapter 2). Likewise, a number of potentially activating fusions were observed that could lead to overexpression of proto-oncogenic kinases (e.g.,

CRKL-MAPK1 and *NRF1-BRAF*; Chapter 3), yet these events were only observed in a single tumor each. Thus, the distribution of driving lesions across prostate tumors may contain a long "tail" of infrequent events that are observed only once or twice among hundreds of tumors. An implication of this model is that large panels of tumors from diverse patient populations will need to be sequenced to identify additional significantly mutated genes. Therefore, ongoing large-scale cancer genome studies will likely prove fruitful for discovering rare driver mutations. In addition, burgeoning efforts to use tumor genomic information to guide clinical decisions may need to overcome the challenge of identifying patient-specific lesions that drive disease but are very rarely recurrent. A similar paradigm has emerged in non-small cell lung cancer. In this disease, oncogenic rearrangements of *ALK* and *ROS1* are proving promising therapeutic targets (Camidge et al., 2012; Janne and Meyerson, 2012); however these lesions occur in roughly 3% and 2% of cases, respectively (Bergethon et al., 2012). Ultimately, both bioinformatics and experimental strategies may be required to identify infrequent, targetable driving lesions within a sufficiently short timeframe to benefit patient care.

Molecular subtypes of prostate cancer and alternative paths of oncogenesis

Our studies add to a growing body of evidence that separable molecular subtypes of prostate cancer exist that harbor distinct sets of co-occuring genetic or epigenetic alterations. For example, we confirmed previous reports *ERG* rearrangement is associated with *TP53* and *PTEN* lesions, as well as deletion or rearrangement of *FOXP1* (Taylor et al., 2010). We observed that *SPOP* mutations are strictly mutually exclusive with ETS fusions, but co-occur with a distinct set of alterations such as *CHD1* deletion, highlighting a distinct subclass of ETS-negative cancer. Previous studies have uncovered ETS-negative subclasses with distinctive transcriptional profiles. For example, roughly 10% of ETS-negative cancers appear to be driven in part by overexpression of the serine protease inhibitor gene *SPINK1* (Tomlins et al., 2008). Similarly, tumors may be classified by expression profiles corresponding to ETS-driven transcription, stem cell-like signatures or activation of one of several other oncogenic pathways. These

classifications appear to portend different survival outcomes (Markert et al., 2011). Therefore, it could be fruitful to establish whether transcriptional profiles of ETS-negative tumors correspond to specific mutational subclasses, as the latter could be easier to assess reliably for clinical applications.

Our analysis of clonal evolution suggests that ETS gene fusions and *SPOP* mutations lead tumors down divergent pathways early in their development, because these alterations are strictly clonal and mutually exclusive. It may be worthwhile, therefore, to study whether these cancer groups are linked to different environmental or genetic risk factors or clinical behaviors. Previous studies comparing clinical outcomes of ETS-positive to ETS-negative tumors have given varying results depending on the populations compared (i.e., PSA-screened or not) and the endpoints assessed (PSA recurrence versus overall mortality) (Demichelis et al., 2007; Rubin et al., 2011; Saramaki et al., 2008). These studies may be aided by further stratification of tumors on a molecular genetic level to account for the genotypic variability of ETS-negative tumors.

A model for transcription-associated DNA damage in ETS-positive prostate tumors

Our analysis of inter-dependent genomic rearrangements suggests that chromosomal abnormalities in prostate cancer may accumulate over a series of relatively few punctuated events (Chapter 3). This process, which we have termed “chromoplexy,” may be driven by different mechanisms in different genetic subclasses of cancers. For example, tumors with *CHD1* disruption displayed an excess of DNA rearrangements and deletions. The breakpoints of these lesions were strongly enriched in gene-poor, late-replicating DNA distant from transcriptionally active chromatin. Our RNA-Seq studies of CHD1-depleted prostate epithelial cells (Chapter 4) suggested that CHD1 suppression induces similar transcriptional changes to genotoxic insults such as gamma irradiation or cisplatin treatment. Thus, *CHD1* disruption might increase the frequency or magnitude of chromoplexy events in a subset of incipient cancer cells, thereby promoting disruption of other cancer genes (Chapter 3) – a conjecture that will be addressed in future experimental studies.

In contrast, chromoplexy in ETS-positive tumors produced distinctive patterns of structural alterations that may indicate a different generative mechanism. Chained rearrangements in ETS-rearranged tumors frequently involved small clusters of breaks distributed across four or more chromosomes. Rearrangements in these tumors were somewhat enriched near actively transcribed DNA. A prior study from our lab indicated that some ETS-rearranged tumors showed enrichment of breaks near androgen receptor (AR) and ERG transcription factor binding sites (Berger et al., 2011), a finding that was replicated in this study.

The association of DNA breaks with active transcription led us to hypothesize that transcription-coupled DNA damage may account for complex rearrangement chains in ETS-positive tumors. Several lines of experimental evidence have indicated that transcription can both engender transient DNA breaks and bring geographically distant foci into close proximity, allowing the formation of specific gene fusions. For example, the *TMPRSS2* and *ERG* loci are brought into physical proximity by liganded AR, and the *TMPRSS2-ERG* fusion can be induced by concomitant irradiation or endonuclease activity of TOP2B, ORF2 or AID (Haffner et al., 2010; Lin et al., 2009; Mani et al., 2009). Similarly, active transcription and physical co-localization precedes the fusion of oncogenes involved in other cancers, such as anaplastic large-cell lymphoma, Burkitt's lymphoma or chronic myeloid leukemia (Robbiani et al., 2008).

Our analysis of prostate cancer genomes suggests that such a process may underlie not only *TMPRSS2-ERG* fusions, but chromoplexy-associated rearrangements in general in ETS-positive tumors. By this model (Figure 5.1), genes throughout the genome that are regulated by the same transcription factor(s) (e.g., AR) are brought into close proximity at "transcription factories" (Cook, 2010). As transcription proceeds, DNA breaks are frequently generated and recognized by double strand break (DSB) repair machinery. Although such DSBs are usually repaired correctly, occasionally two broken DNA ends may be aberrantly joined in an unbalanced translocation. In this case, two DNA ends are left unrepaired and may fuse to additional DSB sites that arise nearby (Figure 5.1). This process could continue within a cell cycle until all broken ends are finally paired, resulting in a "closed chain" of rearrangements. Alternatively, if one or more broken DNA fragments are not re-ligated to a centromere prior to cell division, chains could result that are not "closed." Most instances of chromoplexy would be detrimental to cell survival and proliferation. In rare cases, however, generation of an oncogenic fusion

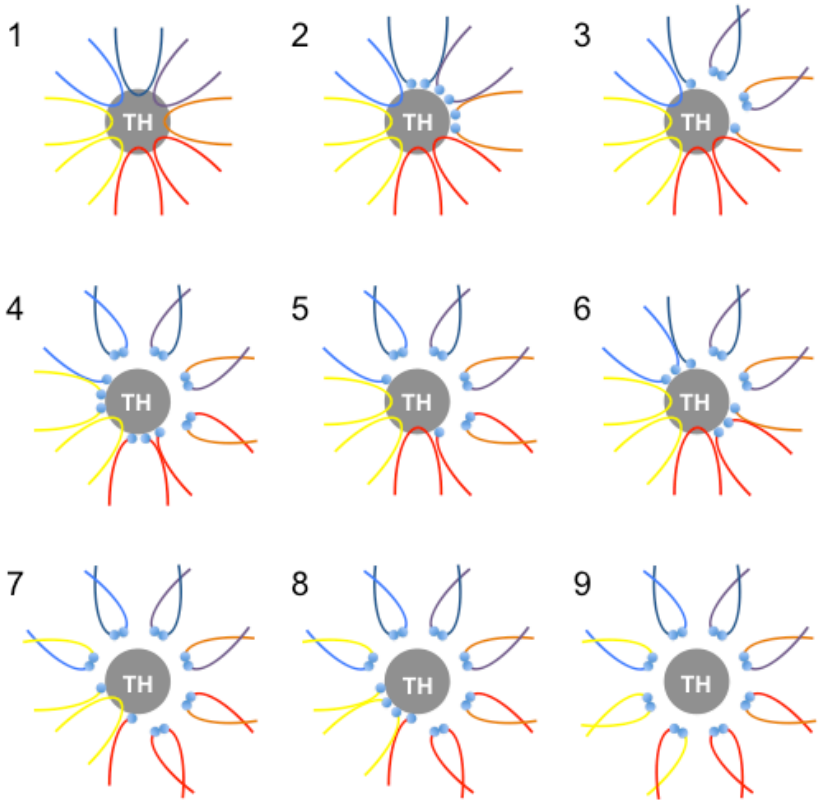


Figure 5.1. Model for chromoplexy via transcription-associated DNA breaks at transcriptional hubs

Closed chains of rearrangement breakpoints associated with chromoplexy could arise from DNA damage associated with transcription. Multiple genomic loci across distinct chromosomes (denoted by strands of different colors) may be brought into physical proximity at transcriptional hubs ("TH", gray circles), for example, due to coordinate regulation by the androgen receptor or other transcription factors (1). Double-strand breaks accumulating at transcription hubs could be recognized by double-strand break repair machinery (2). The aberrant repair of two broken loci in a non-reciprocal manner would leave unpaired DNA ends, that could aberrantly fuse to other DNA ends that arise within the same cell cycle (3). Thus multiple rounds of breaks and rearrangement could be "propagated" until the chain is closed by the last remaining free ends being fused to one another (step 4-9).

such as *TMPRSS2-ERG* or disruption of a tumor-suppressor gene such as *PTEN* could confer a selective advantage that drives the clonal expansion of a cancer cell. This model could be tested by inhibiting DSB repair in prostate epithelial cells in the presence and absence of stimulated AR transcription. The model would predict that chained rearrangements could be detected in the setting of AR activity by FISH or single-cell sequencing. If supported experimentally, this model could point to transcription-associated DSBs as a target for therapy or chemoprevention.

Future directions: prostate cancer in the era of genomics-driven medicine

High-throughput genomic profiling has advanced the understanding, prognostication and treatment of several tumor types. For example, identification of mutations in *BAP1* in uveal melanoma (Harbour et al., 2010) or *IDH1* in glioblastoma and acute myeloid leukemia (Mardis et al., 2009; Parsons et al., 2008) demonstrated the power of genome sequencing to pinpoint novel cancer-driving mutations. Risk-predictive transcriptional signatures have improved prognostication for patients with breast cancer (van 't Veer et al., 2002), while the mutational status of *EGFR* in non-small cell lung cancer predicts clinical response to inhibitors of this kinase (Paez et al., 2004). Prostate cancer may be similarly ripe for discovery of novel cancer genes and biomarkers as well, since genomic characterization of large cohorts of aggressive tumors has only recently become feasible. As a step in this direction, our exome sequencing study revealed that *SPOP* is among the most frequently mutated genes in primary tumors, though its role in cancer was heretofore unrecognized. Ongoing studies will likely identify additional mutations and rearrangements that occur infrequently but nonetheless promote oncogenic growth.

Several hurdles must be overcome for genomic technologies to impact the clinical management of prostate cancer. For instance, biopsies produce scarce material for clinical genotyping and may not fully capture the relevant molecular heterogeneity within a tumor. Expression signatures have not yet demonstrated sufficient prognostic value to merit widespread use. In addition, recurrent genomic lesions identified thus far are largely not considered “druggable”.

These challenges can likely be surmounted by new approaches. For example, genomic characterization may identify opportunities to leverage synthetic lethality by inhibiting targets that are essential in the setting of a particular mutation, such as poly (ADP-ribose) polymerase in ETS-fusion positive prostate cancer (Brenner et al., 2011). The analysis of multiple samples from a primary tumor and perhaps from circulating tumor cells may allow aggressive tumor subclones to be identified. Ultimately, new paradigms for clinical trials may be required that incorporate cancer genomic information. In spite of these challenges, genomic profiling is likely to play an expanding role in the biological study of prostate cancer and eventually in the clinical management of this malignancy.

(Page intentionally left blank)

APPENDIX 1

Supplemental tables

The following tables are provided separately as Excel spreadsheets:

Table S2.2. Exome sequencing statistic summaries

Table S2.3. Somatic mutations in 112 primary prostate tumor-normal pairs

Table S3.1. Clinical characteristics of genome-sequenced prostate tumors

Table S3.2. Sequencing metrics of 57 prostate cancer whole genomes

Table S3.3. Somatic DNA alterations in 57 prostate cancers

Table S3.4. Outlier expression of rearranged genes

Table S3.5. Summary of rearrangement chains

Table S3.6. Clonality analysis of prostate tumor genomes

Table S2.1. Clinical characteristics of exome-sequenced primary prostate cancers

Characteristic	Whole exome-sequenced tumors	
Age, years		
Median (range)	63	(34 – 77)
Pre-operative Serum PSA (ng/μL)		
Median (range)	7.8	(2.7 – 31.5)
Pathologic Stage, N %		
Stage pT3 Total	**44**	**39%**
Stage pT2a	4	4%
Stage pT2b	1	1%
Stage pT2c	39	35%
Stage pT3 Total	**68**	**61%**
Stage pT3a	49	44%
Stage pT3b	18	16%
Stage pT3c	1	1%
Gleason Pattern (major + minor), N %*		
Gleason 3+3	13	12%
Gleason 3+4	58	52%
Gleason 4+3	29	26%
Gleason 4+4	8	7%
Gleason 4+5	4	4%
Percentage of Gleason Pattern 4 and 5, N %*		
0-19%	40	36%
20-39%	23	12%
40-59%	5	5%
60-79%	12	11%
80-100%	29	26%
***TMPRSS2-ERG* Fusion Status, N %** †		
Fusion-negative	53	48%
Fusion with interstitial deletion	34	31%
Fusion without interstitial deletion	24	22%

* Gleason scores based on review of hematoxylin and eosin slides from site of tumor chosen for DNA extraction and sequencing

† *TMPRSS2-ERG* fusion status assessed by fluorescence *in situ* hybridization (FISH)

Table S2.4. Mutation of *PIK3CA* and *PTEN* is enriched in locally advanced tumors.

(**A**) Tumors are grouped based on stage and presence or absence of mutations in *PIK3CA* or *PTEN*. Only stage pT3 tumors displayed mutation in either gene (two-sided $p = 0.011$, Fisher's exact test).

(**B**) Mutations in *PTEN* or *PIK3CA* detected by exome sequencing. The hyper-mutated tumor, PR-00-1165, was included in this analysis and contained two mutations in *PTEN*, presumably affecting separate alleles. Amino acids are numbered based on RefSeq protein ID NP_000305 for *PTEN* and NP_006209 for *PIK3CA*. Instances of all substitutions have been documented previously in prostate cancer or other cancer types in the Cosmic database (http://www.sanger.ac.uk/genetics/CGP/cosmic/). *, nonsense mutation.

A

	Stage pT2	Stage pT3	Total
***PIK3CA/PTEN* mutant:**	0	9	9
***PIK3CA/PTEN* wild-type:**	44	59	103
Total:	44	68	112

B

***PTEN* mutations:**	**Reported in Cosmic**	***PIK3CA* mutations:**	**Reported in Cosmic**
K128N	Yes	p.H1047R	Yes
R130Q	Yes	p.G118D	Yes
Y336*	Yes	p.Q546P	Yes
G129R	Yes	p.Y1021H	Yes
R173H, R233*	Yes, Yes		

Table S2.5. Mutations in significantly-mutated genes and other selected genes

Substitutions in significantly mutated genes are documented at the transcript and protein level for exome-sequenced samples. Annotations are based on the UCSC Genome Browser identifier listed in the left column beneath each gene symbol. The right-most four columns summarize expression data for each gene from the panel of 63 tumors profiled by transcriptome sequencing; "RNA" refers to the $\log_{10}$(RPKM + 1) value for each transcript. All genes except for *SCN11A* and *THSD7B* are expressed at comparatively high levels in prostate tumors at the mRNA level. fs, frame-shift mutation; del, deletion; CV, coefficient of variation.

Table S2.5 (continued)

Gene symbol (Transcript)	Patient	Genomic coordinates (hg19)	Variant type	Codon change	Protein change	Median RNA	Mean RNA	RNA CV	RNA Percentile
SPOP (uc010dbk.2)	P04-2599	g.chr17:47696425A>G	Missense	c.(397-399)TTC>TCC	p.F133S	3.88	3.78	0.13	88%
	P07-684	g.chr17:47696448A>T	Missense	c.(373-375)TTT>TTA	p.F125L				
	P07-837	g.chr17:47696424G>C	Missense	c.(397-399)TTC>TTG	p.F133L				
	P09-649	g.chr17:47696467C>T	Missense	c.(355-357)AGT>AAT	p.S119N				
	PR-00-160	g.chr17:47696421C>G	Missense	c.(400-402)AAG>AAC	p.K134N				
	PR-01-2492	g.chr17:47696425A>C	Missense	c.(397-399)TTC>TGC	p.F133C				
	PR-04-3222	g.chr17:47696424G>T	Missense	c.(397-399)TTC>TTA	p.F133L				
	PR-05-839	g.chr17:47696689A>T	Missense	c.(259-261)TAC>AAC	p.Y87N				
	PR-09-5245	g.chr17:47696432A>C	Missense	c.(391-393)TGG>GGG	p.W131G				
	PR-2661	g.chr17:47696426A>C	Missense	c.(397-399)TTC>GTC	p.F133V				
	PR-2740	g.chr17:47696424G>T	Missense	c.(397-399)TTC>TTA	p.F133L				
	PR-2761	g.chr17:47696425A>G	Missense	c.(397-399)TTC>TCC	p.F133S				
	PR-2915	g.chr17:47696426A>C	Missense	c.(397-399)TTC>GTC	p.F133V				
	PR-3051	g.chr17:47696438T>C	Missense	c.(385-387)AAA>GAA	p.K129E				
TP53 (uc002gim.2)	P03-871	g.chr17:7579417-7579426(del)	Frame shift deletion	c.(259-270)fs	p.P87fs	3.99	3.95	0.16	89%
	P04-1243	g.chr17:7577121G>A	Missense	c.(817-819)CGT>TGT	p.R273C				
	P06-4428	g.chr17:7577547C>T	Missense	c.(733-735)GGC>GAC	p.G245D				
	P06-4428	g.chr17:7578457C>T	Missense	c.(472-474)CGC>CAC	p.R158H				
	P07-5036	g.chr17:7577538C>T	Missense	c.(742-744)CGG>CAG	p.R248Q				
	PR-02-254	g.chr17:7578542G>A	Missense	c.(388-390)CTC>TTC	p.L130F				
	PR-0415	g.chr17:7577529A>T	Missense	c.(751-753)ATC>AAC	p.I251N				
	PR-06-1999	g.chr17:7579359G>A	Missense	c.(328-330)CGT>TGT	p.R110C				

Table S2.5 (continued)

Gene symbol (Transcript)	Patient	Genomic coordinates (hg19)	Variant type	Codon change	Protein change	Median RNA	Mean RNA	RNA CV	RNA Percentile
PTEN (uc001kfb.2)	P04-1790	g.chr10:89692900G>C	Missense	c.(382-384)AAG>AAC	p.K128N	2.78	2.55	0.29	75%
	P06-3676	g.chr10:89692905G>A	Missense	c.(388-390)CGA>CAA	p.R130Q				
	PR-02-254	g.chr10:89720857C>G	Nonsense	c.(1006-1008)TAC>TAG	p.Y336*				
	PR-02-254	g.chr10:89692901G>C	Missense	c.(385-387)GGA>CGA	p.G129R				
	PR-3026	g.chr10:89711874G>A	Splice site mutation	N/A	p.G165_splice				
FOXA1 (uc001wuf.2)	P05-620	g.chr14:38061231A>T	Missense	c.(757-759)ATG>AAG	p.M253K	6.92	6.94	0.09	99%
	PR-04-1367	g.chr14:38061294G>A	Missense	c.(694-696)GCA>GTA	p.A232V				
	PR-0427	g.chr14:38061231A>C	Missense	c.(757-759)ATG>AGG	p.M253R				
	PR-09-3687	g.chr14:38061313C>T	Missense	c.(676-678)GAC>AAC	p.D226N				
CDKN1B (uc001rat.2)	P00-000450	g.chr12:12870910A>G	Missense	c.(136-138)GAG>GGG	p.E46G	3.90	3.72	0.15	88%
	P03-1426	g.chr12:12871788-12871788(del)	Frame shift deletion	c.(505-507)fs	p.R169fs				
	PR-00-160	g.chr12:12871228-12871228(del)	Frame shift deletion	c.(454-456)fs	p.R152fs				
ZNF595 (uc003fzv.1)	P03-1334	g.chr4:59330T>C	Missense	c.(10-12)GTA>GCA	p.V4A	1.23	1.25	0.36	53%
	P04-594	g.chr4:59357T>A	Missense	c.(37-39)TTC>TAC	p.F13Y				
	PR-09-3687	g.chr4:85888A>T	Missense	c.(493-495)ATA>TTA	p.I165L				
	PR-09-5702	g.chr4:86116C>A	Missense	c.(721-723)ACA>AAA	p.T241K				
THSD7B (uc002tva.1)	PR-01-1934	g.chr2:138163182G>C	Splice site mutation	N/A	p.A803_splice	0.07	0.13	1.41	27%
	PR-03-1026	g.chr2:137988713C>T	Missense	c.(1729-1731)ACG>ATG	p.T577M				
	PR-05-839	g.chr2:138000058C>T	Missense	c.(2089-2091)CGC>TGC	p.R697C				
	PR-09-2767	g.chr2:138373832G>C	Missense	c.(3424-3426)GTG>CTG	p.V1142L				
	PR-09-5702	g.chr2:138033591C>A	Missense	c.(2401-2403)ACA>AAA	p.T801K				
	PR-2872	g.chr2:137928505C>T	Missense	c.(1627-1629)CGC>TGC	p.R543C				

Table S2.5 (continued)

Gene symbol (Transcript)	Patient	Genomic coordinates (hg19)	Variant type	Codon change	Protein change	Median RNA	Mean RNA	RNA CV	RNA Percentile
MED12 (uc004dyy.2)	P04-594	g.chrX:70349258C>T	Missense	c.(3670-3672)CTC>TTC	p.L1224F	2.92	2.93	0.12	77%
	P07-684	g.chrX:70349946C>A	Missense	c.(3928-3930)CCA>CAA	p.P1310Q				
	PR-00-160	g.chrX:70344951C>A	Missense	c.(2179-2181)GAC>GAA	p.D727E				
	PR-3026	g.chrX:70349258C>T	Missense	c.(3670-3672)CTC>TTC	p.L1224F				
	PR-3035	g.chrX:70349258C>T	Missense	c.(3670-3672)CTC>TTC	p.L1224F				
	PR-3036	g.chrX:70349257T>G	Missense	c.(3667-3669)GTT>GTG	p.V1223V				
NIPA2 (uc001yux.2)	P03-1906	g.chr15:23012398A>C	Missense	c.(328-330)CTT>CGT	p.L110R	3.58	3.61	0.17	85%
	P07-837	g.chr15:23006670C>A	Missense	c.(634-636)GTG>TTG	p.V212L				
	PR-09-3421	g.chr15:23014493C>A	Missense	c.(232-234)GCG>TCG	p.A78S				
PIK3CA (uc003fjk.2)	PR-04-639	g.chr3:178952085A>G	Missense	c.(3139-3141)CAT>CGT	p.H1047R	1.39	1.31	0.28	55%
	PR-09-2517	g.chr3:178917478G>A	Missense	c.(352-354)GGT>GAT	p.G118D				
	PR-09-2744	g.chr3:178936095A>C	Missense	c.(1636-1638)CAG>CCG	p.Q546P				
	PR-09-3421	g.chr3:178952006T>C	Missense	c.(3061-3063)TAC>CAC	p.Y1021H				
C14orf49 (uc001yei.3)	P00-000450	g.chr14:95884213C>T	Missense	c.(2878-2880)GCC>ACC	p.A960T	0.38	0.42	0.57	38%
	P07-718	g.chr14:95932380C>T	Missense	c.(514-516)CGG>CAG	p.R172Q				
	PR-09-5700	g.chr14:95899740C>A	Nonsense	c.(2545-2547)GAG>TAG	p.E849*				
	PR-2916	g.chr14:95921763G>A	Missense	c.(1087-1089)GCG>GTG	p.A363V				
	PR-3023	g.chr14:95906394G>A	Missense	c.(1930-1932)CGG>TGG	p.R644W				
SCN11A (uc011ays.1)	P07-144	g.chr3:38926826A>T	Nonsense	c.(3016-3018)TTA>TAA	p.L1006*	0.02	0.03	1.15	24%
	PR-00-1823	g.chr3:38938566G>A	Missense	c.(2173-2175)CGT>TGT	p.R725C				
	PR-01-2554	g.chr3:38936068C>A	Nonsense	c.(2791-2793)GAA>TAA	p.E931*				
	PR-06-1999	g.chr3:38938451C>T	Missense	c.(2287-2289)CGC>CAC	p.R763H				
	PR-09-2517	g.chr3:38938451C>T	Missense	c.(2287-2289)CGC>CAC	p.R763H				

Table S2.5 (continued)

Category	Gene symbol (Annotation transcript)	Patient	Genomic coordinates (hg19)	Variant type	Codon change	Protein change	Median RNA	Mean RNA	RNA CV	RNA Percentile	Refs.
Genes mutated at sites reported and experimentally studied in other cancer types	***IDH1*** (uc002vcs.2)	P06-2325	g.chr2:209113112C>T	Missense	c.(394-396)CGT>CAT	p.R132H	6.00	5.94	0.14	98%	(Dang, 2009); (Parsons, 2008)
		PR-05-3595	g.chr2:209113112C>T	Missense	c.(394-396)CGT>CAT	p.R132H					
	AKT1 (uc001ypk.2)	P06-1125	g.chr14:105246551C>T	Missense	c.(49-51)GAG>AAG	p.E17K	5.51	5.51	0.06	97%	(Carpten, 2007)
	HRAS (uc001lpv.2)	PR-09-5702	g.chr11:533874T>C	Missense	c.(181-183)CAG>CGG	p.Q61R	4.98	4.94	0.08	95%	(Carter, 1990) (Buhrman 2007)
	ATM (uc001pkb.1)	P00-000450	g.chr11:108218029G>T	Missense	c.(8608-8610)GAT>TAT	p.D2870Y	1.33	1.30	0.30	54%	(Canman, 1998); (Greenman, 2007); (Sasaki, 1998)
		PR-09-2767	g.chr11:108216537C>T	Missense	c.(8485-8487)CCA>CTA	p.P2829L					
		PR-2682	g.chr11:108201148-108201148(del)	Frame shift deletion	c.(7513-7515)fs	p.K2505fs					

Table S2.5 (continued)

Category	Gene symbol (Annotation transcript)	Patient	Genomic coordinates (hg19)	Variant type	Codon change	Protein change	Median RNA	Mean RNA	RNA CV	RNA Percentile	Refs.
Genes with damaging mutation that undergo recurrent damaging mutations in other cancer types	***APC*** (uc010jby.2)	P05-620	g.chr5:112173635A>T	Nonsense	c.(2344-2346)AAG>TAG	p.K782*	1.11	1.18	0.34	51%	(Miyaki, 1994); (Beroud, 1996)
		PR-00-160	g.chr5:112157644-112157648(del)	Frame shift deletion	c.(1363-1368)fs	p.K455fs					
		PR-0415	g.chr5:112175979-112175979(del)	Frame shift deletion	c.(4687-4689)fs	p.L1563fs					
	PIK3R1 (uc003jva.2)	PR-09-3566	g.chr5:67591246A>G	Splice site mutation	N/A	p.M582_splice	2.29	2.22	0.23	69%	(Parsons, 2010)
	MLL2 (uc001rta.3)	PR-2661	g.chr12:49438067G>A	Nonsense	c.(5104-5106)CGA>TGA	p.R1702*	3.26	3.13	0.17	81%	(Pasqualucci, 2011); (Parsons, 2010)
	MLL3 (uc003wla.2)	PR-3048	g.chr7:151859711C>A	Nonsense	c.(10951-10953)GAG>TAG	p.E3651*	2.25	2.22	0.21	68%	(Pasqualucci, 2011); (Parsons, 2010)
	KDM6A (uc011mkz.1)	P07-360	g.chrX: 44941865-44941872(del)	Frame shift deletion	c.(3343-3354)fs	p.S1115fs	2.00	2.01	0.18	64%	(van Haaften, 2009)
		PR-2916	g.chrX:44945155G>T	Missense	c.(3634-3636)GGC>GTC	p.G1212V					
		PR-3036	g.chrX:44945146T>A	Missense	c.(3625-3627)ATA>AAA	p.I1209K					
	ARID1A (uc001bmv.1)	P04-2599	g.chr1:27101666C>T	Nonsense	c.(4948-4950)CAG>TAG	p.Q1650*	4.03	3.98	0.09	89%	(Wiegand, 2010)
		PR-09-3421	g.chr1:27099857A>T	Missense	c.(3736-3738)ATG>TTG	p.M1246L					
	EPHA7 (uc003poe.2)	P04-1084	g.chr6:93967815A>C	Splice site mutation	N/A	p.G704_splice	0.92	0.95	0.47	48%	(Guan, 2009)
	NCOR1 (uc002gpo.2)	P05-620	g.chr17:16012098A>C	Splice site mutation	N/A	p.V728_splice	3.07	3.10	0.13	79%	(Gui, 2011)
		P02-2480	g.chr17:15935762G>A	Missense	c.(7171-7173)CGG>TGG	p.R2391W					
	MLL (uc001ptb.2)	PR-2682	g.chr11:118355577G>T	Nonsense	c.(4219-4221)GAG>TAG	p.E1407*	2.35	2.31	0.16	69%	(Gui, 2011)

Table S2.6. Significantly mutated gene sets

Curated gene sets representing canonical pathways were analyzed for enrichment of mutations in their constituent genes (SOM). Sets that are significantly mutated above a q-value of 0.05 (Benjamini-Hochberg adjustment) are listed.

Gene Set	No. mutations	No. tumors with a mutation	B-H q-value	Genes in set mutated (no. tumors affected)
SA_G1_AND_S_PHASES	11	10	0.00033	CDKN1B(3), TP53(8)
RBPATHWAY	14	13	0.00033	ATM(3), CDC25A(1), MYT1(1), TP53(8), WEE1(1)
P53HYPOXIAPATHWAY	14	13	0.0048	ABCB1(2), AKT1(1), ATM(3), TP53(8)
TERTPATHWAY	9	8	0.0048	SP1(1), TP53(8)
PLK3PATHWAY	11	10	0.0062	ATM(3), TP53(8)
IGF1MTORPATHWAY	14	13	0.0072	AKT1(1), GSK3B(1), IGF1R(1), INPPL1(1), PIK3CA(4), PIK3R1(1), PTEN(5)
ARFPATHWAY	13	12	0.008	PIK3CA(4), PIK3R1(1), TP53(8)
G1PATHWAY	16	14	0.008	ATM(3), CDC25A(1), CDKN1B(3), GSK3B(1), TP53(8)
G2PATHWAY	17	16	0.011	ATM(3), CDC25A(1), EP300(1), MYT1(1), PRKDC(1), RPS6KA1(1), TP53(8), WEE1(1)
CHEMICALPATHWAY	16	14	0.011	AKT1(1), ATM(3), CASP3(1), CASP9(1), TLN1(2), TP53(8)
PTENPATHWAY	17	16	0.011	AKT1(1), BCAR1(1), CDKN1B(3), ILK(1), PIK3CA(4), PIK3R1(1), PTEN(5), SHC1(1)
RNAPATHWAY	9	8	0.011	DNAJC3(1), TP53(8)
P53PATHWAY	11	10	0.017	ATM(3), TP53(8)
COMPLEMENT_ACTIVATION_CLASSICAL	10	10	0.017	C1QB(1), C1S(1), C3(3), C6(1), C8A(1), C9(2), MASP1(1)
CLASSICPATHWAY	9	9	0.018	C1QB(1), C1S(1), C3(3), C6(1), C8A(1), C9(2)
COMPPATHWAY	10	10	0.018	C1QB(1), C1S(1), C3(3), C6(1), C8A(1), C9(2), MASP1(1)
HCMVPATHWAY	10	10	0.018	AKT1(1), CREB1(1), MAP2K6(1), MAPK14(1), PIK3CA(4), PIK3R1(1), SP1(1)
TELPATHWAY	14	13	0.025	AKT1(1), EGFR(1), IGF1R(1), POLR2A(1), TEP1(1), TP53(8), XRCC5(1)
ALTERNATIVEPATHWAY	7	7	0.027	C3(3), C6(1), C8A(1), C9(2)
CDC42RACPATHWAY	8	7	0.027	ACTR2(1), PDGFRA(2), PIK3CA(4), PIK3R1(1)
SA_PTEN_PATHWAY	12	11	0.04	AKT1(1), ILK(1), PIK3CA(4), PTEN(5), SHC1(1)
RACCYCDPATHWAY	11	11	0.04	AKT1(1), CDKN1B(3), HRAS(1), PIK3CA(4), PIK3R1(1), RAF1(1)
CELL_CYCLE_KEGG	29	28	0.042	ATM(3), BUB3(1), CDC20(1), CDC25A(1), CDC6(1), CDH1(1), EP300(1), ESPL1(1), GSK3B(1), HDAC2(2), HDAC3(1), HDAC5(2), MAD1L1(1), MCM4(1), PRKDC(1), SMAD4(1), TP53(8), WEE1(1)
IGF1RPATHWAY	11	11	0.042	AKT1(1), HRAS(1), IGF1R(1), IRS1(1), PIK3CA(4), PIK3R1(1), RAF1(1), SHC1(1)

Table S2.7. *SPOP* mutations in multiple cohorts

Cohort	*SPOP* mutation prevalence	Technology	Mutated residues
WCMC	13.3 % (11/83)	WES, RNA-seq, Sanger	Y87N, Y87C, F102C, K129E, F133V, F133S, F133L, F133C
Uropath	10.1% (9/89)	WES	Y87N, S119N, F125L, W131G, F133S, F133L, F133C, K134N
UM	6.1% (3/49)	RNA-seq	F102C, F133L, F133V
UHZ	8.3% (16/193)	Sanger	Y87N, F102C, F102S, W131C, F133V, F133L
UW	14.5% (6/39)	Sanger	F102C, F102D, W131G, F133V

Table S2.8. Somatic copy number alterations associated with *SPOP* mutation

					Association with *SPOP* mutation:					
					Discovery set		Validation set		Combined	
Type	Cytoband	Area Coordinates (hg18)	Peak Coordinates (hg18)	q.values	P-value	Odds ratio	P-value	Odds ratio	P-value	Odds ratio
Del	6q15	chr6:84993071-90693659	chr6:89844069-89854517	7.60E-14	0	49.0936	0.0494	4.6152	0	14.9182
Del	5q14.3	chr5:86736848-92949317	chr5:90698354-90718446	4.76E-06	0	28.9982	0.0023	10.5654	0	18.5816
Del	6q21	chr6:102623781-118338352	chr6:107452805-107484929	2.72E-12	0	30.1055	0.0003	14.5979	0	23.0189
Del	5q21.1	chr5:96540739-99903527	chr5:98211239-98298151	6.24E-17	0	26.1837	0	23.6731	0	26.5255
Del	5q21.1	chr5:98288994-100177733	chr5:99892827-99956161	1.12E-11	0.0001	13.0406	0.0006	11.2936	0	12.8012
Del	6q14.1	chr6:79992719-89377471	chr6:82929067-83023893	3.50E-10	0.0001	10.9692	0.0427	4.4101	0	7.6868
Del	5q21.3	chr5:100266177-130524037	chr5:108694667-108775892	2.46E-06	0.0012	7.412	0.0001	20.5115	0	10.8653
Del	6q13	chr6:71718724-79638241	chr6:74188319-74228252	6.62E-05	0.0017	6.8441	0.2702	2.4264	0.0015	4.4487
Del	21q22.3	chr21:41569692-42033506	chr21:41754509-41805040	3.20E-13	0.0043	0	0.1836	0	0.0004	0
Del	5q11.2	chr5:54500232-57785740	chr5:55428291-55450265	3.84E-08	0.0044	5.4502	0.0107	7.6381	0.0004	4.9509
Amp	8q24.3	chr8:142130394-143948850	chr8:143649163-143686346	0.005918	0.0104	4.7801	0.4292	0.3188	0.1905	1.9013
Del	17p13.1	chr17:7099211-7923319	chr17:7723851-7759444	2.24E-07	0.0157	0.1101	0.6699	0.409	0.0089	0.1662
Del	2q21.1	chr2:114430192-140707944	chr2:131673017-131764875	0.0027275	0.0179	5.1272	0.0191	6.0238	0.0007	5.6709
Del	3p13	chr3:70098695-73514455	chr3:71884403-71891255	2.46E-06	0.0188	0	0.6731	0.4708	0.0179	0.1198
Amp	21q21.1	chr21:18980067-18997817	chr21:18980170-18994744	0.024377	0.0279	7.9899	1	0	0.144	2.9523
Amp	3q22.3	chr3:138503860-138525897	chr3:138503876-138514781	0.0046072	0.0284	3.8842	1	1.0223	0.1342	2.3893
Amp	7q11.23	chr7:71946015-71962929	chr7:71948417-71962425	0.076891	0.0333	4.1148	0.6154	1.7416	0.0554	2.9006
Amp	7q21.2	chr7:91212265-91215184	chr7:91212282-91215163	0.059607	0.0333	4.1148	0.6154	1.7416	0.0554	2.9006
Del	2q23.3	chr2:151939541-152050232	chr2:151970139-152045679	0.074786	0.0349	4.8599	0.65	1.4242	0.0829	2.9745
Del	17q21.31	chr17:39091591-40194083	chr17:39820952-39945190	1.17E-05	0.0366	0	0.1836	0	0.0026	0
Amp	7p15.2	chr7:26355665-26394413	chr7:26356137-26393733	0.059607	0.0435	3.7332	0.5879	2.2144	0.0554	2.9006

Table S2.9. Primer sequences

Amplification	**sense (5' --> 3')**	**antisense (5' --> 3')**
SPOP		
Exons 6 and 7	TTCTATGGGGCCTGCATTT	CTCCACTTGGGGCTTTTTCT
Sequencing	**sense (5' --> 3')**	**antisense (5' --> 3')**
SPOP		
Exon 6	TTTTCTATCTGTTTTGGACAGG	CAAAGCCACAACTTGTCAGTG
Exon 7	TTTGCGAGTAAACCCCAAAG	CTCATCAGATCTGGGAACTGC
qPCR	**sense (5' --> 3')**	**antisense (5' --> 3')**
SPOP	CTTCTGCGAGGTGAGTGTTG	TCCCACAGTCCTCCTAACTCA
GAPDH	TGCACCACCAACTGCTTAGC	GGCATGGACTGTGGTCATGAG

Table S2.10. BAC probes used for FISH

Assay	BAC
***ERG* break-apart**	
5' probe	RP11-372O17
3' probe	RP11-24A11
***ETV1* break-apart**	
5' probe	RP11-661L15
3' probe	RP11-79G16
***ETV4* break-apart**	
5' probe	RP11-147C10
3' probe	CTD-3215I16
***ETV5* break-apart**	
5' probe	RP11-822O23
3' probe	RP11-480B15
***PTEN* deletion**	
PTEN probe	CTD-2047N14
Reference probe	RP11-431P18

APPENDIX 2

Supplemental figures

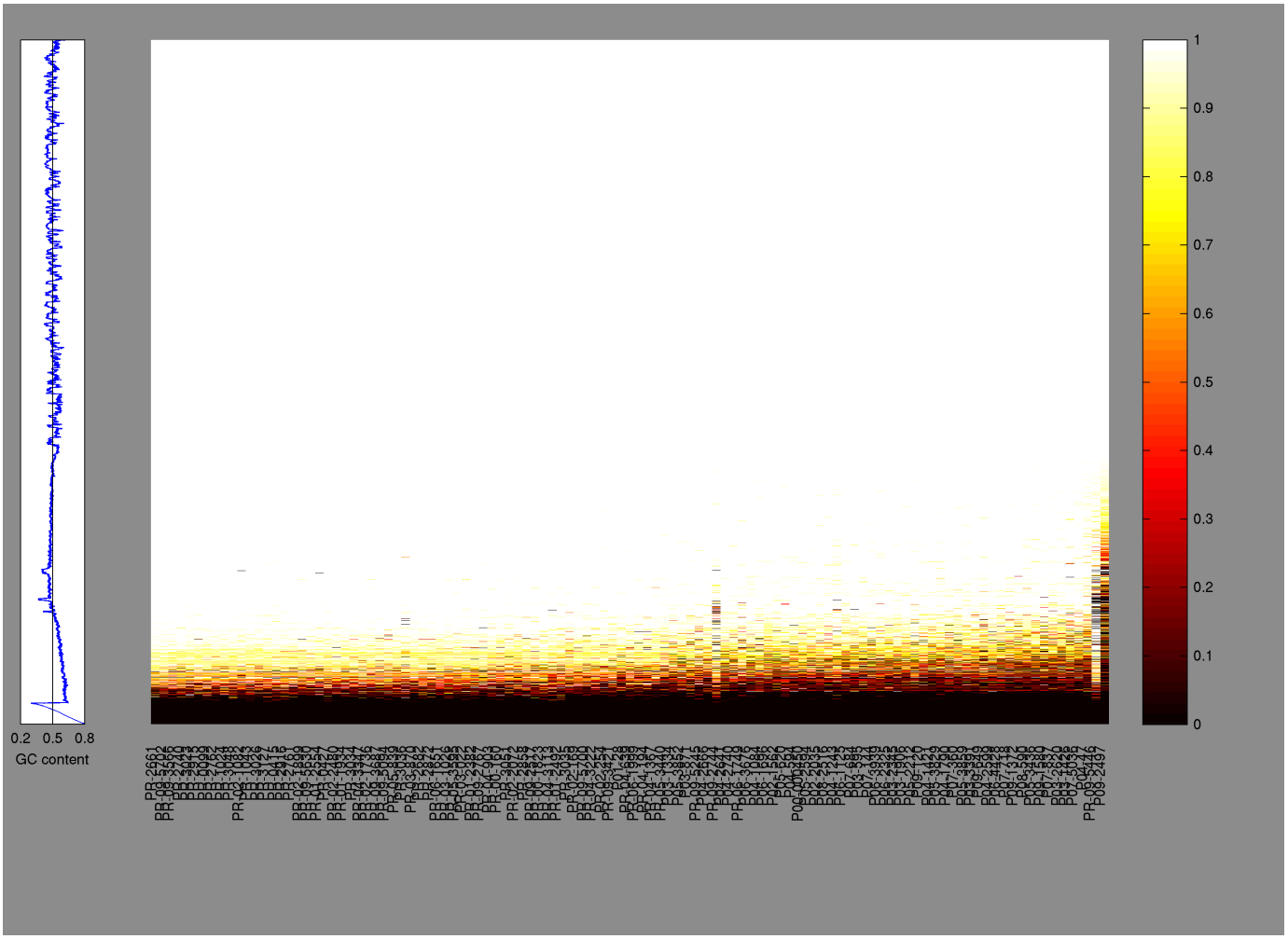


Figure S2.1. Depth and breadth of exome sequencing coverage

(**Center**) Sequencing coverage across all sites targeted by hybrid capture. Each row represents a targeted exonic site; each column represents a tumor-normal pair. Coloring reflects the depth of sequencing coverage. White coloring indicates a minimum of 14 reads in the tumor and 8 reads in the normal.

(**Left**) GC content across targeted sites (GC content is equal to the number of C or G nucleotides divided by the total number of nucleotides).

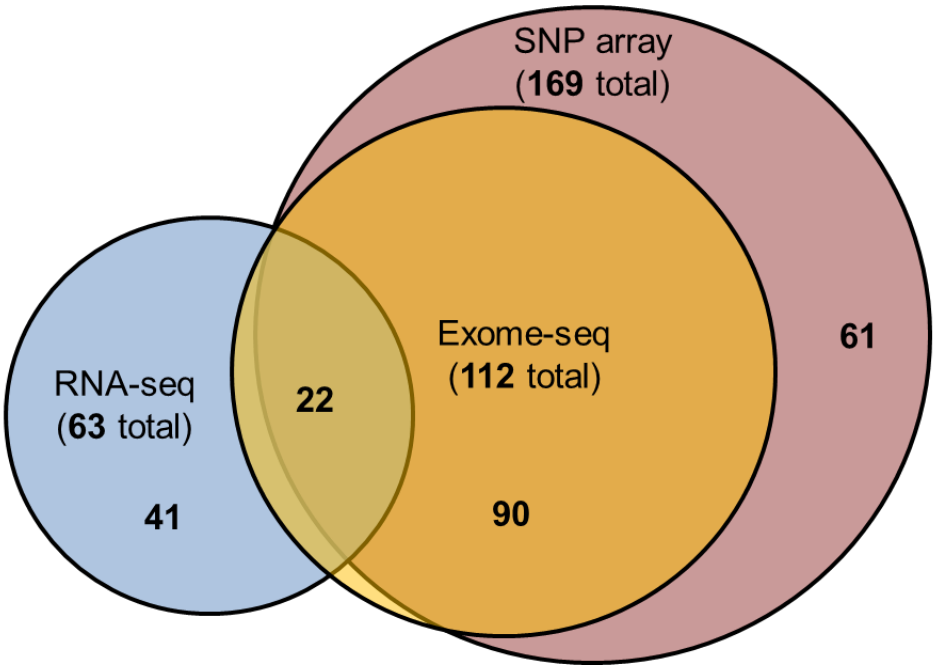


Figure S2.2. Overlap of sample profiling across platforms

Exome sequencing was conducted on 112 tumor-normal pairs. A single highly-mutated tumor (PR-00-1165) was excluded from subsequent analyses, except where otherwise indicated, leaving 111 pairs. RNA-sequencing was performed on 22 of the exome-sequenced tumors and 41 independent tumors. All but four of the 112 exome-sequenced tumors, plus an additional 61 tumors, were analyzed for copy number alteration by high-density SNP array (169 total).

Figure S2.3. Rates of somatic substitutions in prostate exomes

(**A**) Number of somatic mutations per Mb sequenced across the cohort of tumors. A single primary tumor (PR-00-1165) harbored a large excess of mutations compared to other tumors (32.1 per Mb versus 1.4 per Mb median in the remaining primary tumors, red). A prostate cancer metastasis sequenced but not reported here showed a similar extent of mutation (PR-18248; 29.0 mutations per Mb, red). The two highly mutated tumors contained the indicated alterations in mismatch repair genes.

(**B**) Median number of non-synonymous and synonymous mutations across 111 exomes (the single hyper-mutated primary tumor PR-00-1165, with 997 mutations, is excluded).

(**C**) Mutations per million sites sequenced for the most frequent mutation categories in the dataset. *CpG to T, C to T transversion at a CpG dinucleotide; *CpG to A/C/T, C to T transversion not in the context of a CpG dinucleotide; C to (G/A), mutation of C to G or A; A to mut, mutation of A; Indel, small insertion or deletion.

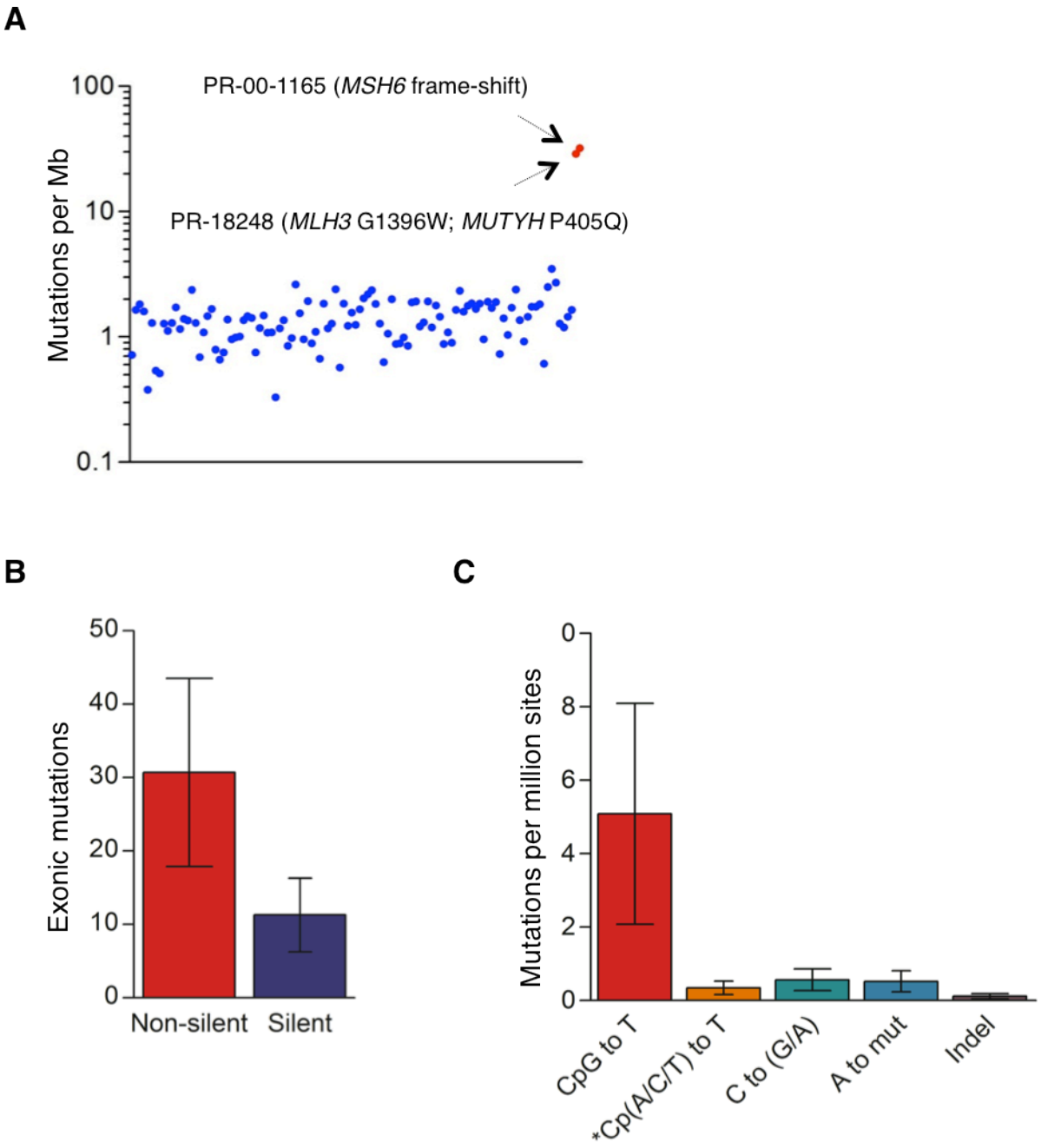


Figure S2.3 (continued)

Figure S2.4. Relative ability to detect mutations in subgroups of tumors

(**A-D**) The allelic fraction (AF) values of mutations were used to assess the relative purity of cancer DNA in each tumor (SOM). AF is the number of reads supporting a mutant allele divided by the total number of reads covering the mutated site. (A) and (B), maximum mutant AF observed in each tumor, grouped by Gleason score and stage. The top fifth percentile of AF values was removed in each tumor to exclude values that were elevated due to copy number variation at mutated sites. (C) and (D), as in (A) and (B), but showing median AF values across all mutations for each tumor. The relative purity of cancer DNA as assessed by AF did not vary by pathological stage or Gleason score.

(**E**) Maximum mutant AF correlated only moderately with the number of mutations detected, implying that detection of mutations in most tumors was not systematically limited by admixture of normal DNA.

(**F**) Mutations per Mb sequenced in tumors grouped by source of paired normal DNA: peripheral blood (n=22) or non-cancerous prostate (n=89). No difference in mutation rates was observed between the two groups. Two-tailed p-values were calculated using the Mann Whitney test (two groups) or Kruskal-Wallis (multiple groups) test and the Spearman test for correlation. n.s., not significant. Statistical analysis was performed using GraphPad Prism.

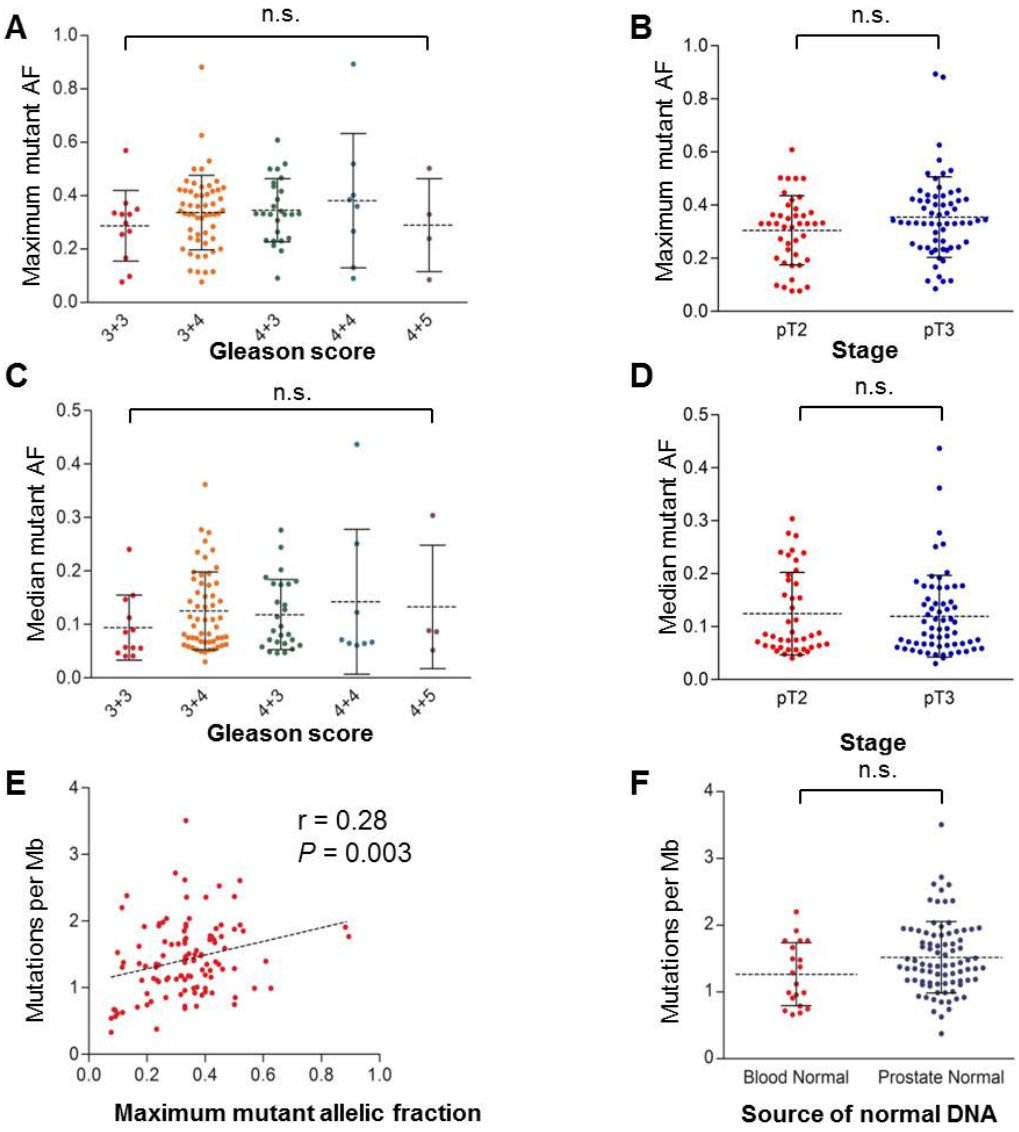


Figure S2.4 (continued)

Figure S2.5. Mutational landscape across a spectrum of primary prostate cancers

(**A-D**) Mutations per Mb of covered DNA sequence for 111 primary prostatic adenocarcinomas grouped by clinical parameters. The horizontal axes denote: (**A**) Gleason score (major pattern + minor pattern); (**B**) Percent of cancer with Gleason pattern 4 or 5 histology; (**C**) Pathological tumor stage, where T3 indicates extra-prostatic extension; (**D**) Presence or absence of the *TMPRSS2-ERG* fusion based on fluorescence in-situ hybridization (FISH). Mutation rates are higher in pT3 tumors but do not vary by Gleason pattern or *TMPRSS2-ERG* fusion status.

(**E**) Fraction of mutations in each tumor that are C to T transitions at CpG sites.

(**F**) Number of CpG to T transitions per million CpG sites. Both the number and proportion of CpG to T mutations is increased in *TMPRSS2-ERG* fusion positive tumors. Two-tailed p-values from the Mann Whitney test (two groups) or Kruskal-Wallis test (multiple groups) are indicated for each comparison. Statistical analysis was performed using GraphPad Prism. n.s., not significant.

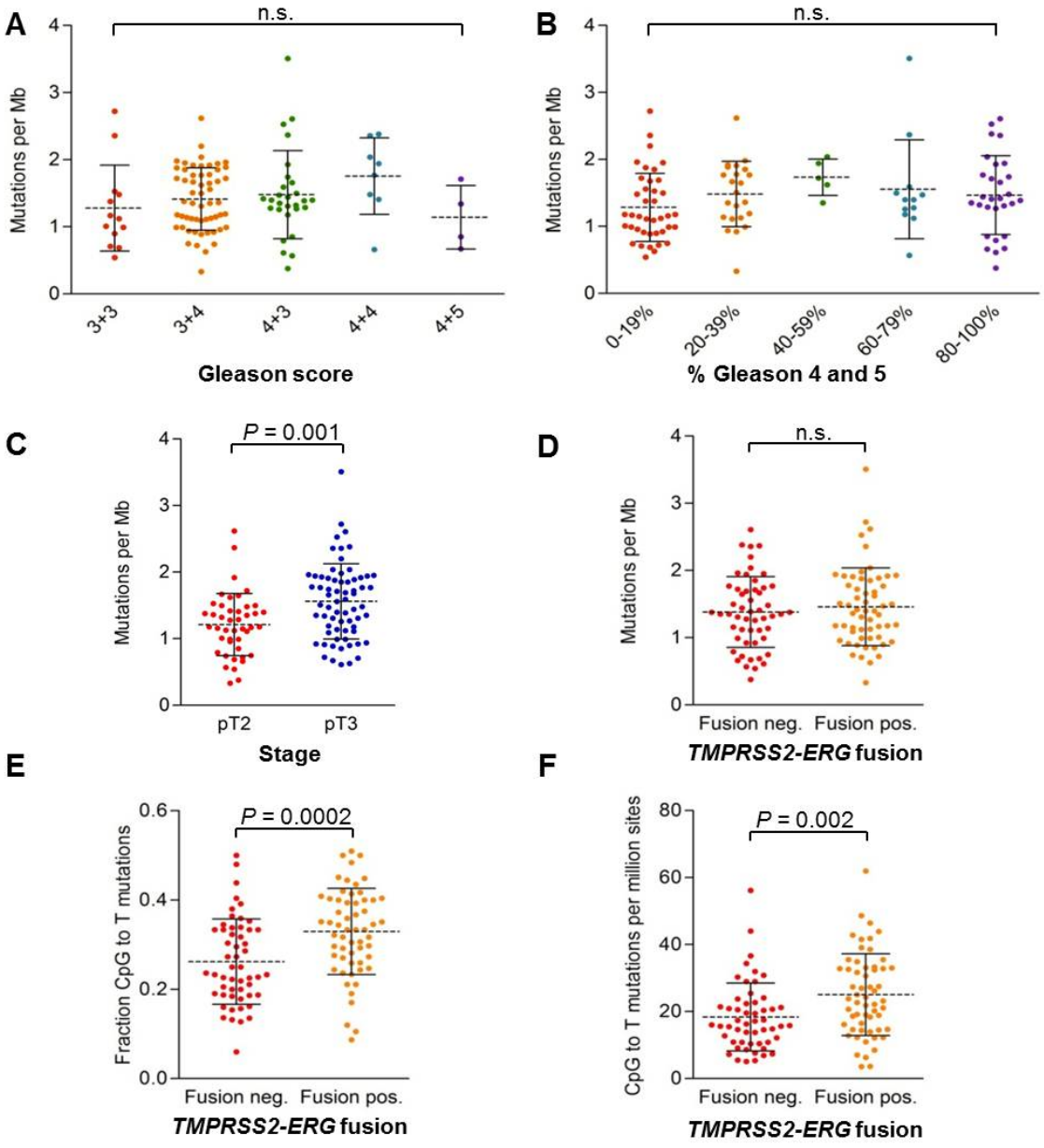


Figure S2.5 (continued)

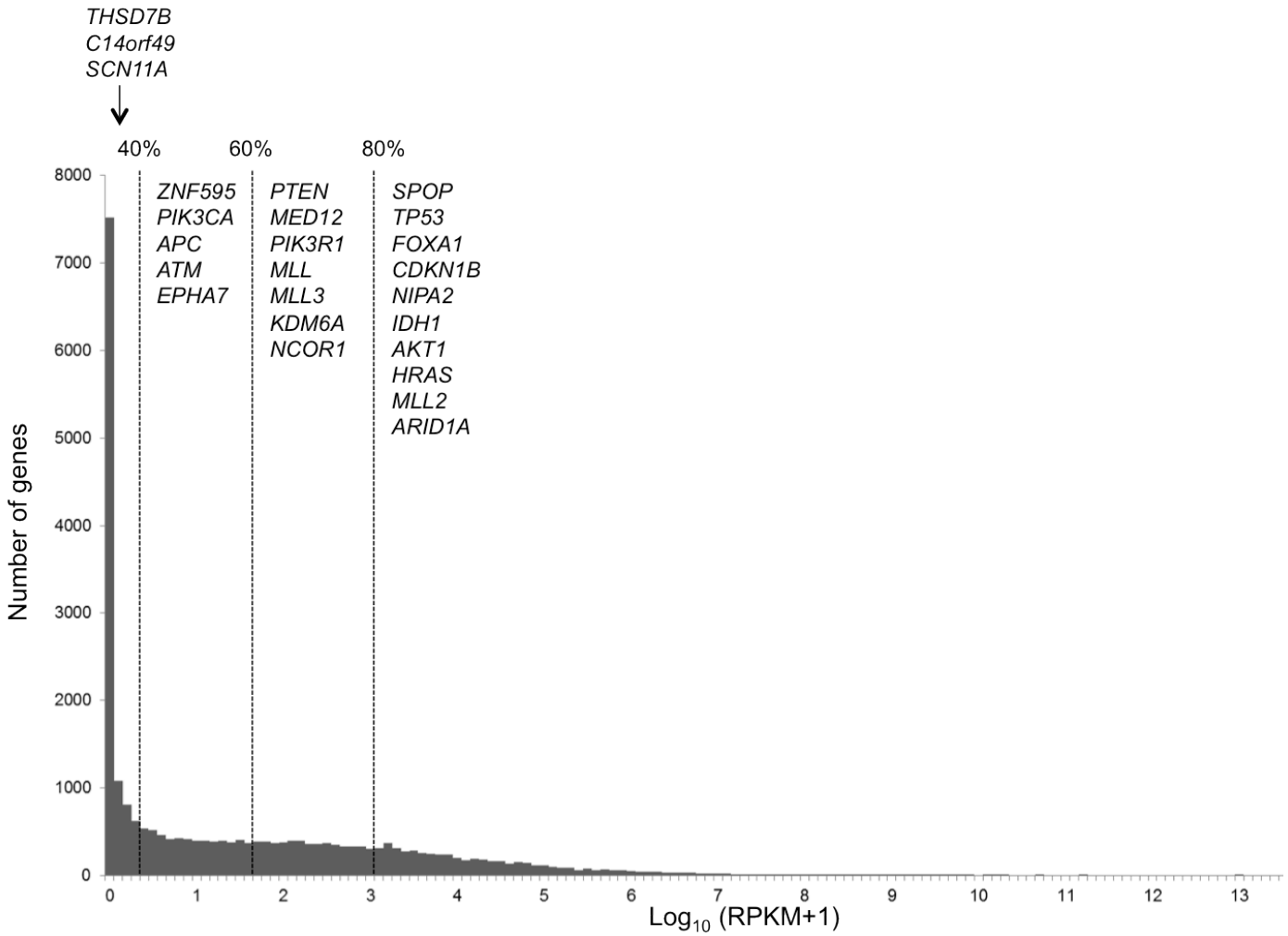


Figure S2.6. Expression levels of select mutated genes

Significantly-mutated genes and selected genes listed in Table S2.5 were analyzed for level of transcript expression in the RNA-Seq dataset. The histogram shows the number of transcripts with a given value of $\log_{10}$ (RPKM+1) (where RPKM is the number of reads per kilobase of exon per million mapped sequence reads), binned by increments of 0.1. The RPKM provides an estimate of the relative expression of transcripts. Vertical lines indicate the percentile of $\log_{10}$(RPKM+1) among all transcripts. Listed genes are grouped based on their percentile of $\log_{10}$(RPKM+1) value: <40%, 40-60%, 60-80% and >80%. Values and percentiles are listed in Table S2.5.

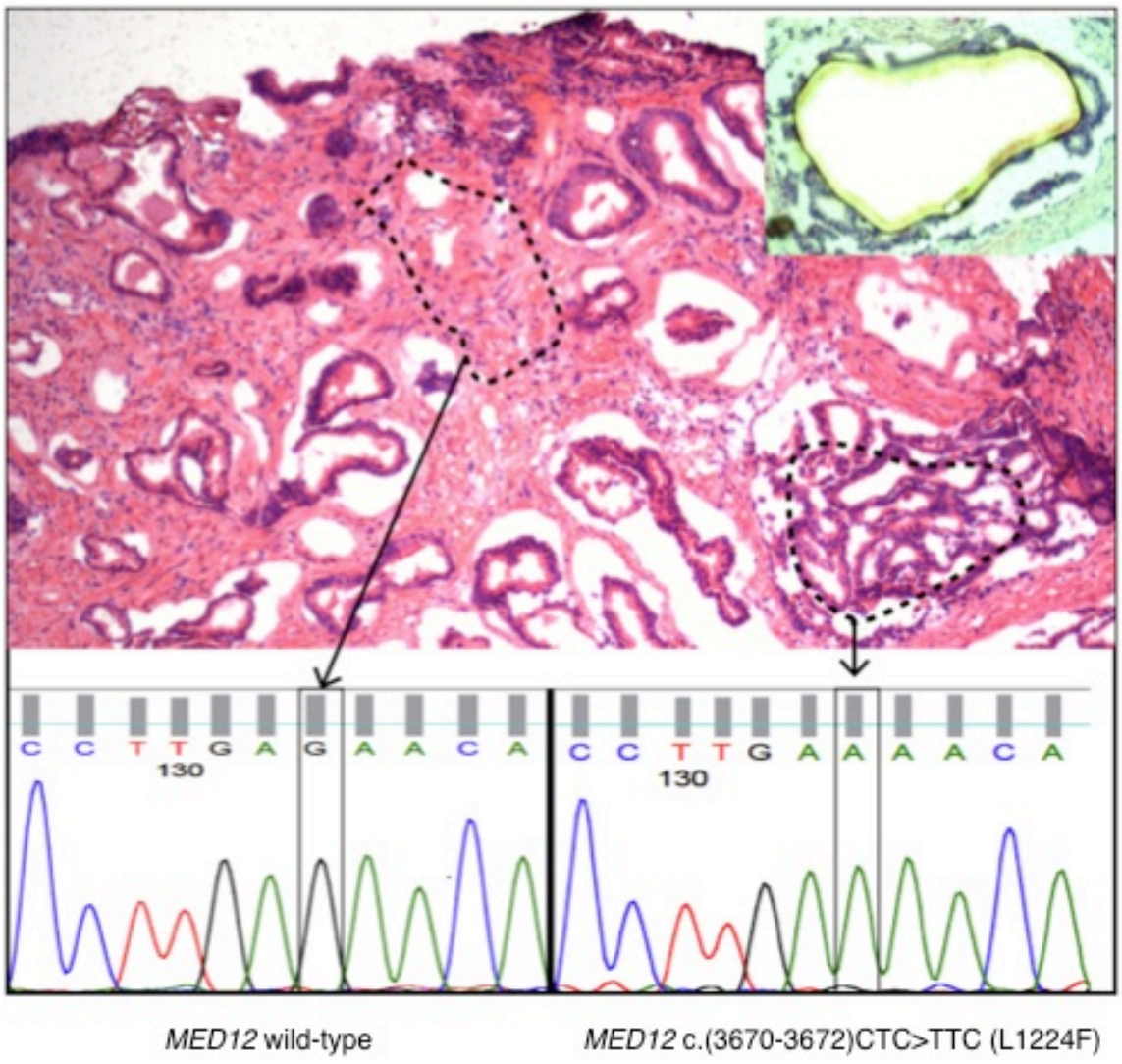


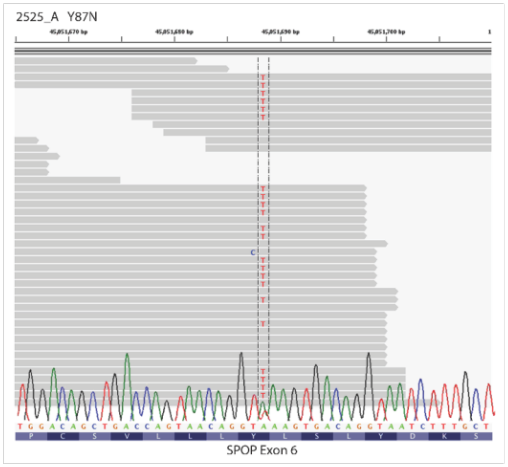
Figure S2.7. Laser capture microdissection and sequencing of *MED12*

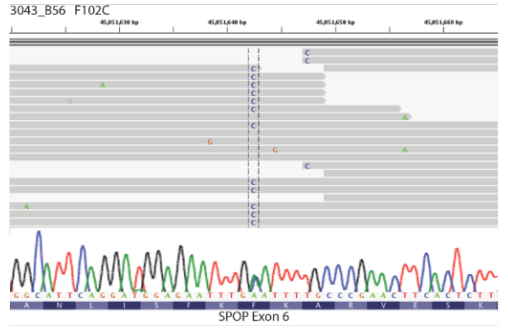
(**Top**) Laser capture micro-dissection and Sanger sequencing was performed on *MED12*-mutant tumors to determine whether the mutations were present in epithelial or stromal cells. H&E slide of frozen tissue from a *MED12*-mutant tumor (PR-3026) showing adenocarcinoma and surrounding mixed stroma. Exome sequence reads demonstrated an L1224F mutation in exon 26 of *MED12*. Laser capture micro-dissection was performed to separate epithelium from stroma (inset).

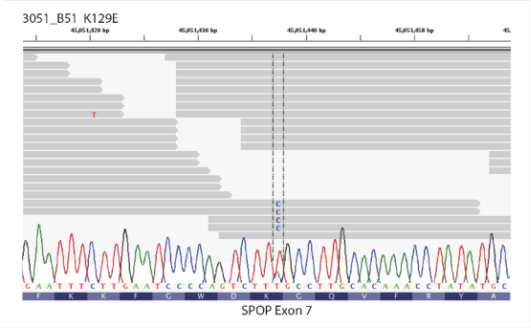
(**Bottom**) The selected stromal area (dashed line, left) demonstrates wild-type *MED12* sequence by Sanger sequencing, while the dissected tumor gland (dashed line, right) exhibits the L1224F mutation.

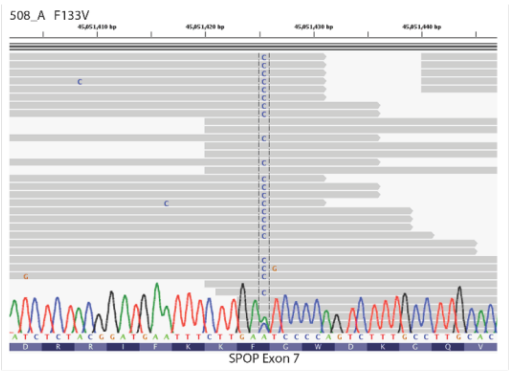
Figure S2.8. Mutations in *SPOP* in RNA-Seq data and Sanger sequencing of genomic tumor DNA

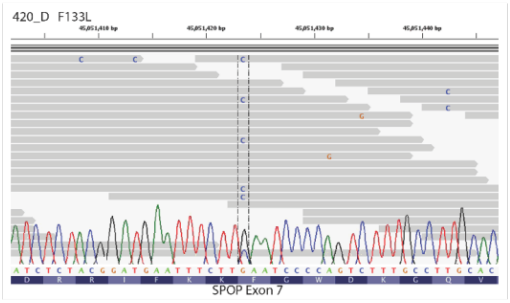
In each panel, RNA-seq reads mapping to SPOP Exon 6 or 7 from the indicated sample are shown. Coordinates (hg18) on chromosome 17 are at the top of each panel, and the reference genome (hg18) and wild-type SPOP amino acids are displayed at the bottom. Each horizontal gray bar represents one read. Nucleotide mismatches with respect to the reference genome in each read are highlighted by displaying the mismatched base. The Sanger tracing of genomic DNA from the same tumor focus is overlayed below the RNA-seq reads.

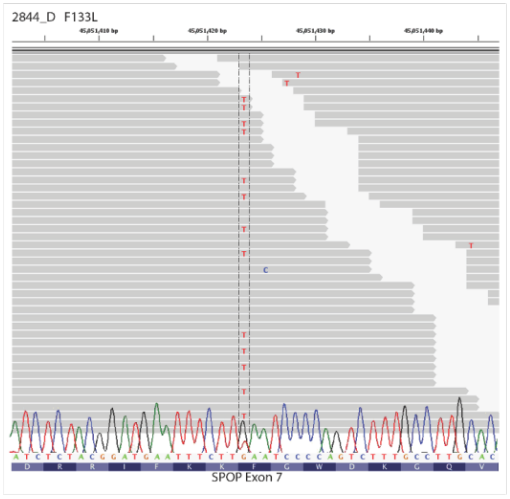












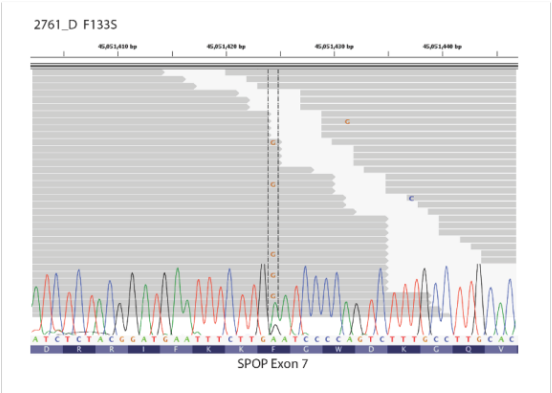


Figure S2.8 (continued)

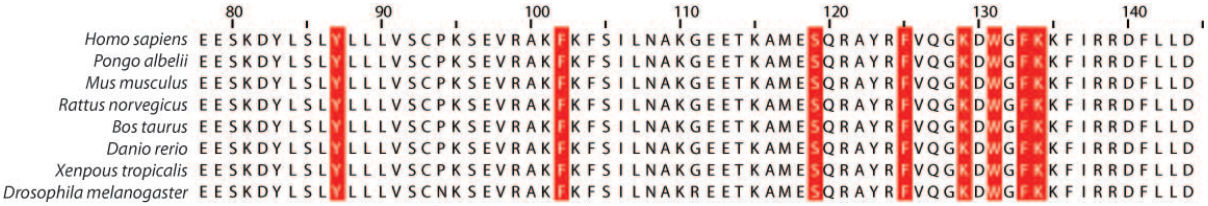


Figure S2.9. Multiple sequence alignment of the MATH domain of SPOP across species

Multiple sequence alignment was performed with ClustalW2 and visualized using Jalview. Residues mutated in prostate cancer (Y87, F102, S119, F125, K129, W131, F133, K134) are highlighted.

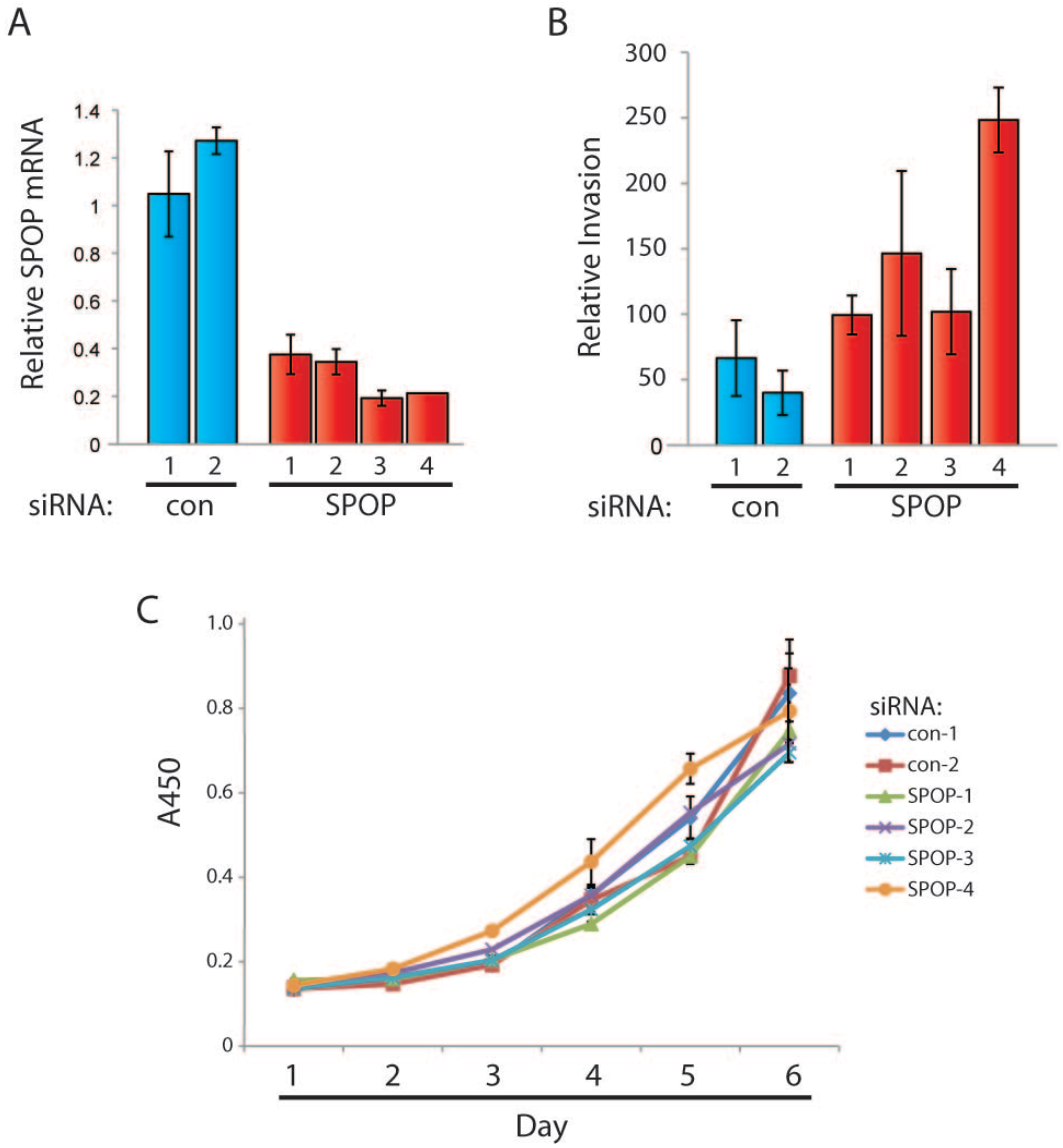


Figure S2.10. Multiple independent siRNAs targeting *SPOP* have similar effects in prostate cell lines

(**A**) Expression of *SPOP* mRNA in DU145 cells transfected with 2 different control siRNAs and 4 different *SPOP* siRNAs, normalized to *GAPDH* expression, by real-time RT-PCR.

(**B**) Quantitation of invaded DU145 cells transfected with control and SPOP siRNAs in Matrigel invasion assays.

(**C**) Growth curves of DU145 cells transfected with control and *SPOP* siRNAs, measured with WST-1 assay.

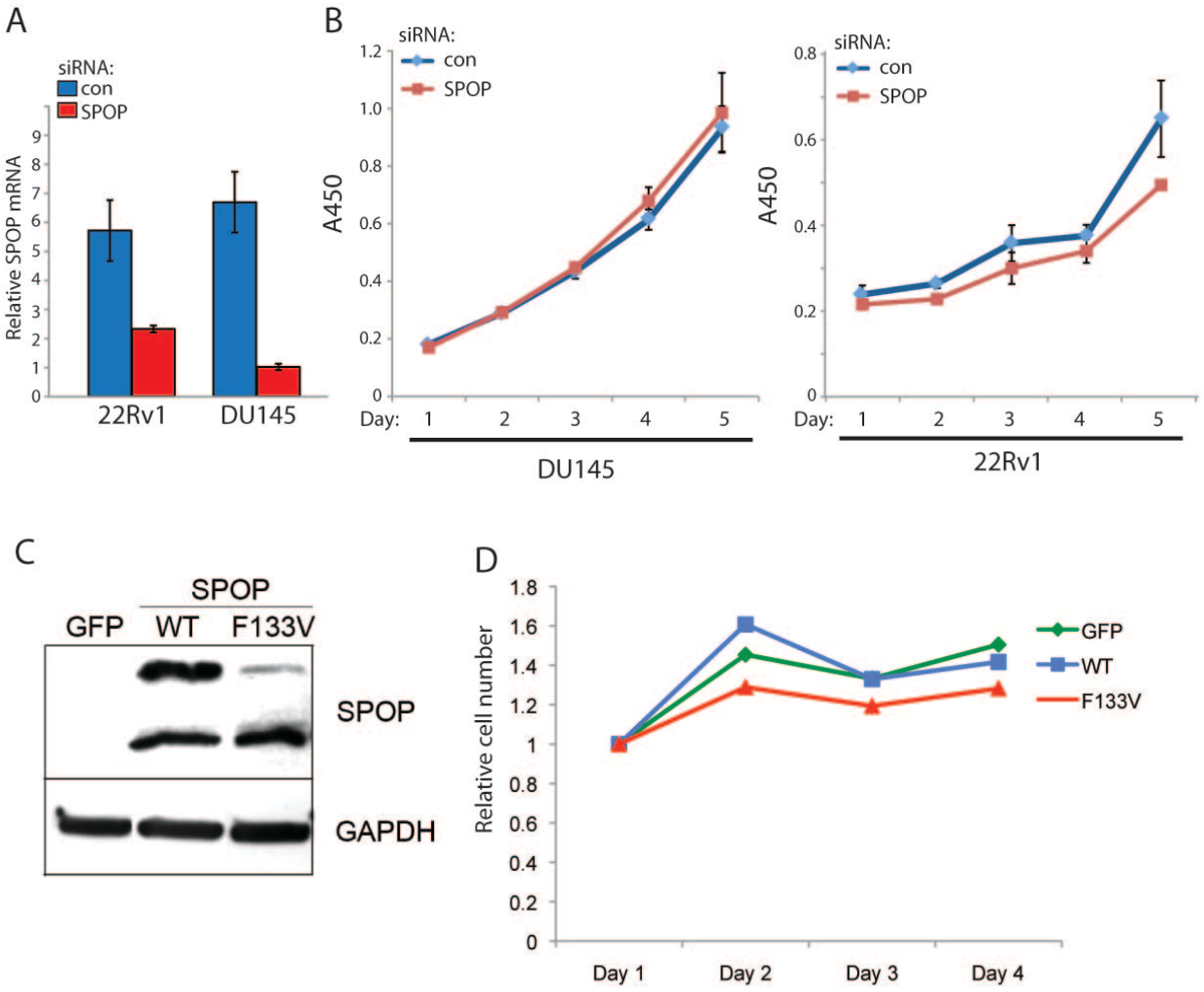


Figure S2.11. Transfection with SPOP siRNA or SPOP mutant does not affect cell growth or viability

(**A**) Expression of *SPOP* mRNA in 22Rv1 and DU145 cells transfected with control and SPOP siRNA, normalized to GAPDH expression, by real-time RT-PCR.

(**B**) Growth curves of 22Rv1 and DU145 cells transfected with control and SPOP siRNA, measured with WST-1 assay.

(**C**) Western blot showing SPOP expression in DU145 cells transfected with SPOP wild-type and F133V.

(**D**) Growth curves of DU145 cells transfected with SPOP wild-type and F133V, measured with WST-1 assay.

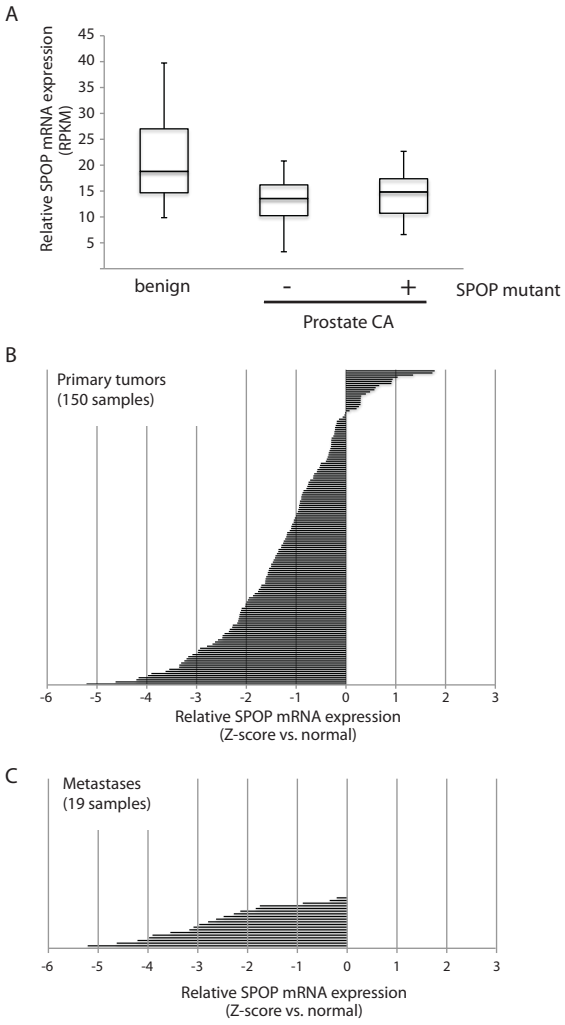


Figure S2.12. *SPOP* is not upregulated in prostate cancer.

(**A**) *SPOP* mRNA expression measured by RNA-seq in 6 benign prostate samples and 53 prostate cancers (7 *SPOP* mutant, 46 *SPOP* wild-type). Relative expression is displayed as reads per kilobase per million mapped reads (RPKM).

(**B**-**C**) *SPOP* mRNA expression from a publicly available dataset (www.cbioportal.org/cgx/)[12] in 150 primary prostate cancers (B) and 19 metastases (C). Relative expression is displayed as Z-score versus matched normal; positive = increased expression, negative = decreased expression.

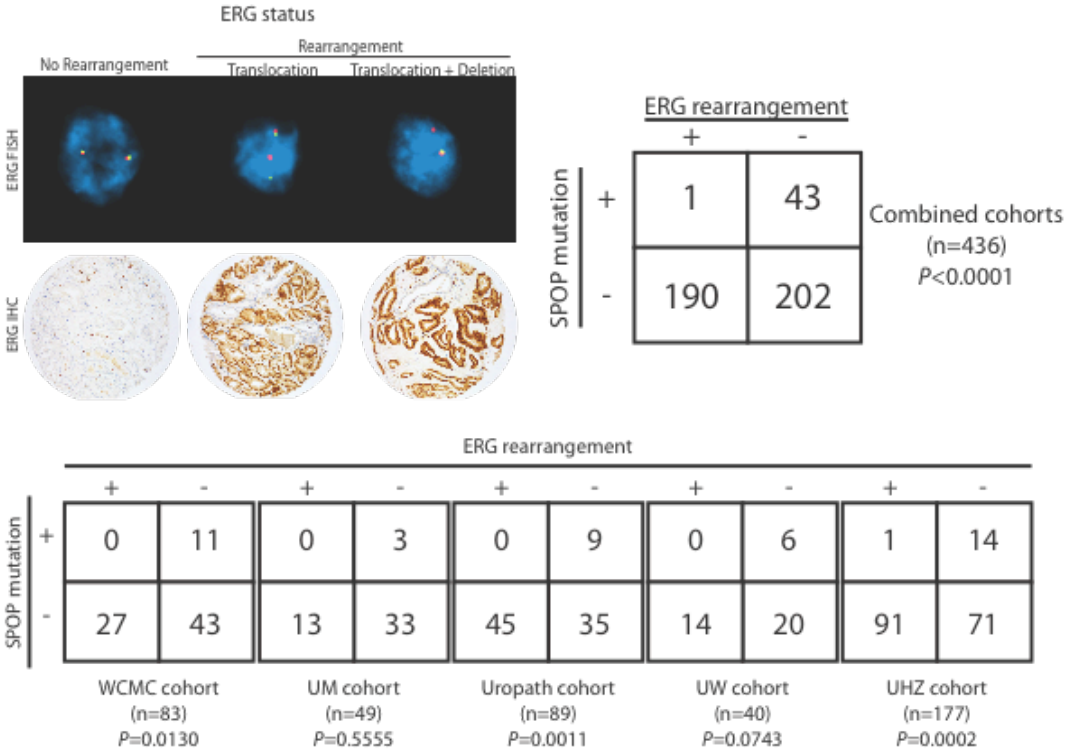


Figure S2.13. Tumors with *SPOP* mutation lack *ETS* rearrangements

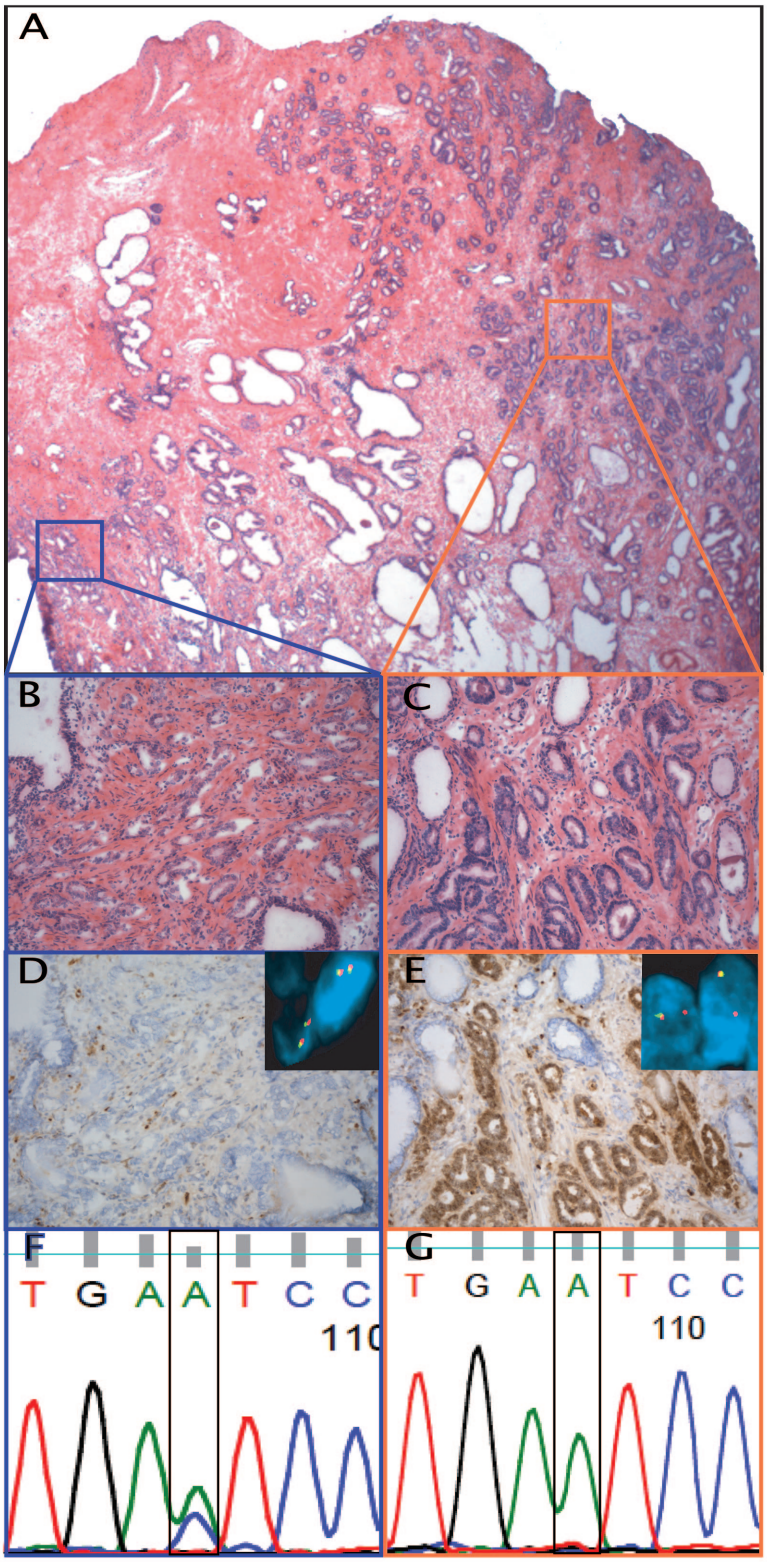
Relationship of *SPOP* mutation and *ERG* rearrangement. *ERG* rearrangement was determined by FISH and IHC.

Figure S2.14. Separate foci of prostate adenocarcinoma with mutually exclusive ERG-rearranged and SPOP-mutated status

(**A**) Low power view of two distinct foci of prostate adenocarcinoma in a prostatectomy specimen (H&E stained slide of frozen tissue, original magnification 2x).

(**B**) The tumor on the left side (blue box) has Gleason score 3+4=7, is ERG-negative by immunohistochemistry (D) without *ERG* rearrangement by FISH (inset), and demonstrates the F133V *SPOP*-mutation (F).

(**C**) The tumor on the right side (orange box) has Gleason score 3+3=6, is ERG-positive by immunohistochemistry (E) with *ERG*-rearrangement by FISH (inset), and demonstrates *SPOP* wild-type sequence (G).



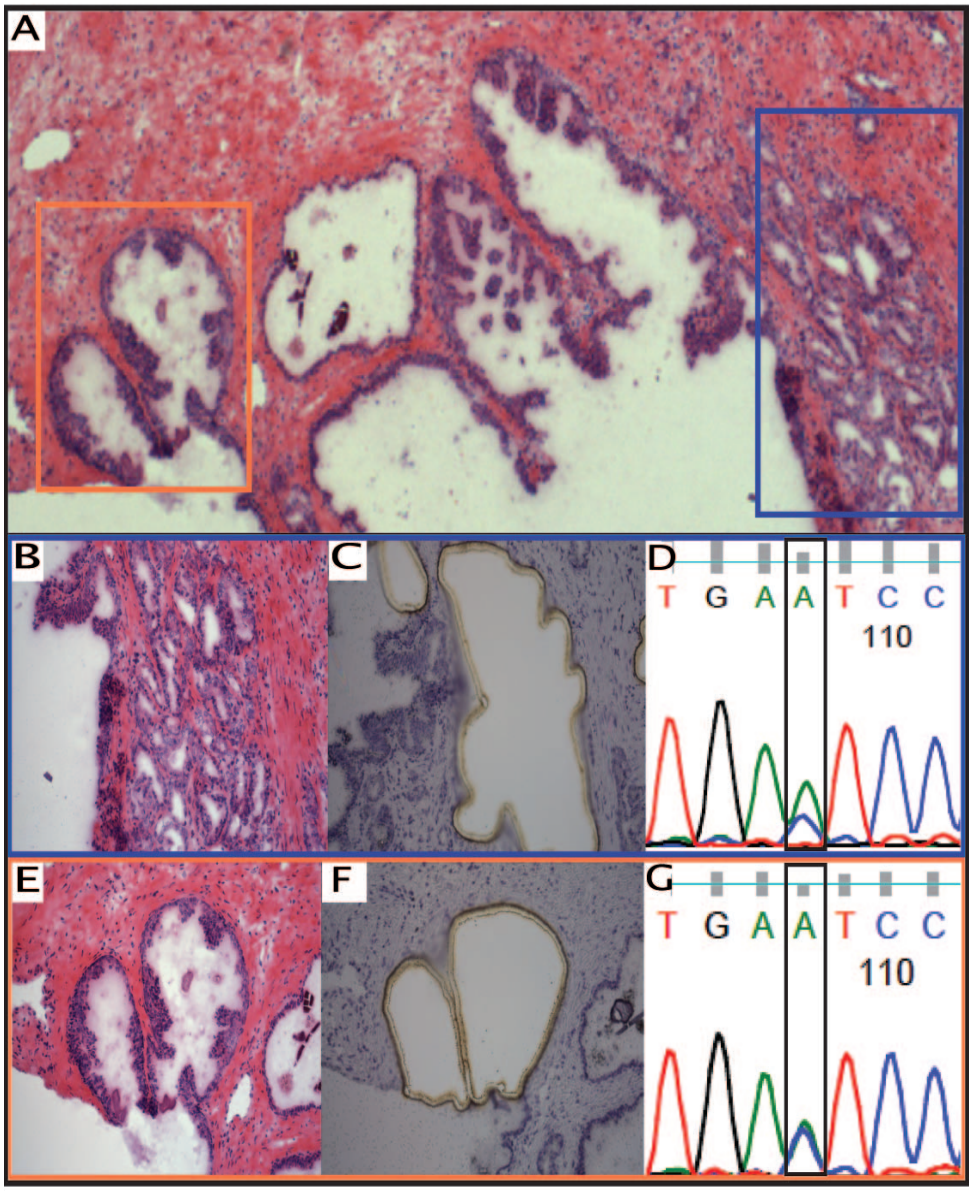


Figure S2.15. Detection of SPOP mutation in high-grade prostatic intraepithelial neoplasia (HGPIN)

(**A**) Low power view of prostate adenocarcinoma (blue box) and HGPIN (orange box) in a prostatectomy specimen.

(**B**-**G**) Cancer area before (B) and after (C) Laser Capture Microdissection (LCM). Images of HGPIN before (E) and after (F) LCM. DNA sequence demonstrates F133V *SPOP*-mutation in both adenocarcinoma (D) and HGPIN (G).

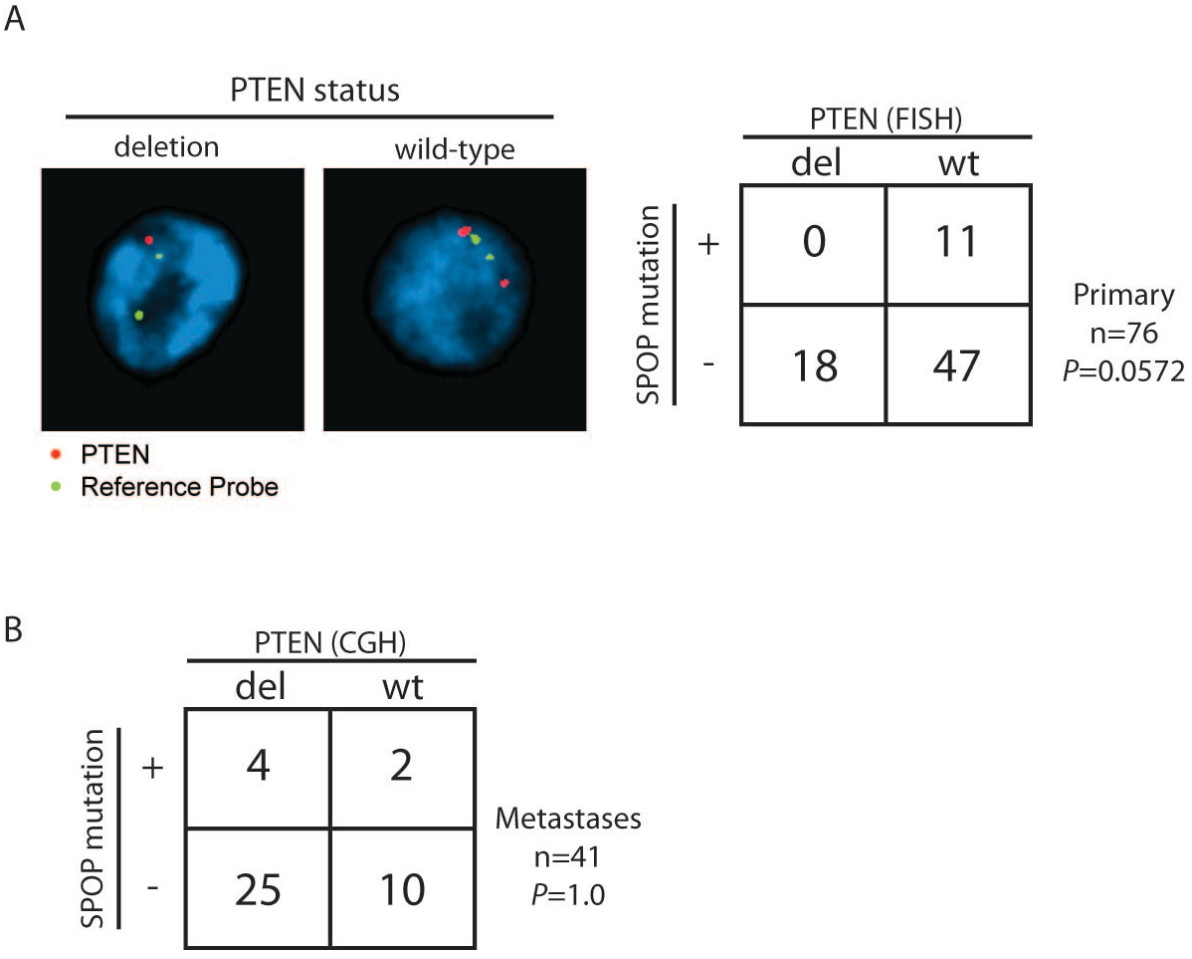


Figure S2.16. Tumors with *SPOP* mutation lack *PTEN* deletion in primary but not metastatic prostate cancer

(**A**) Relationship of *SPOP* mutation and *PTEN* deletion determined by FISH in primary prostate cancers from the WCMC cohort.

(**B**) Relationship of *SPOP* mutation and *PTEN* deletion determined by CGH in prostate cancer metastases from the UW cohort.

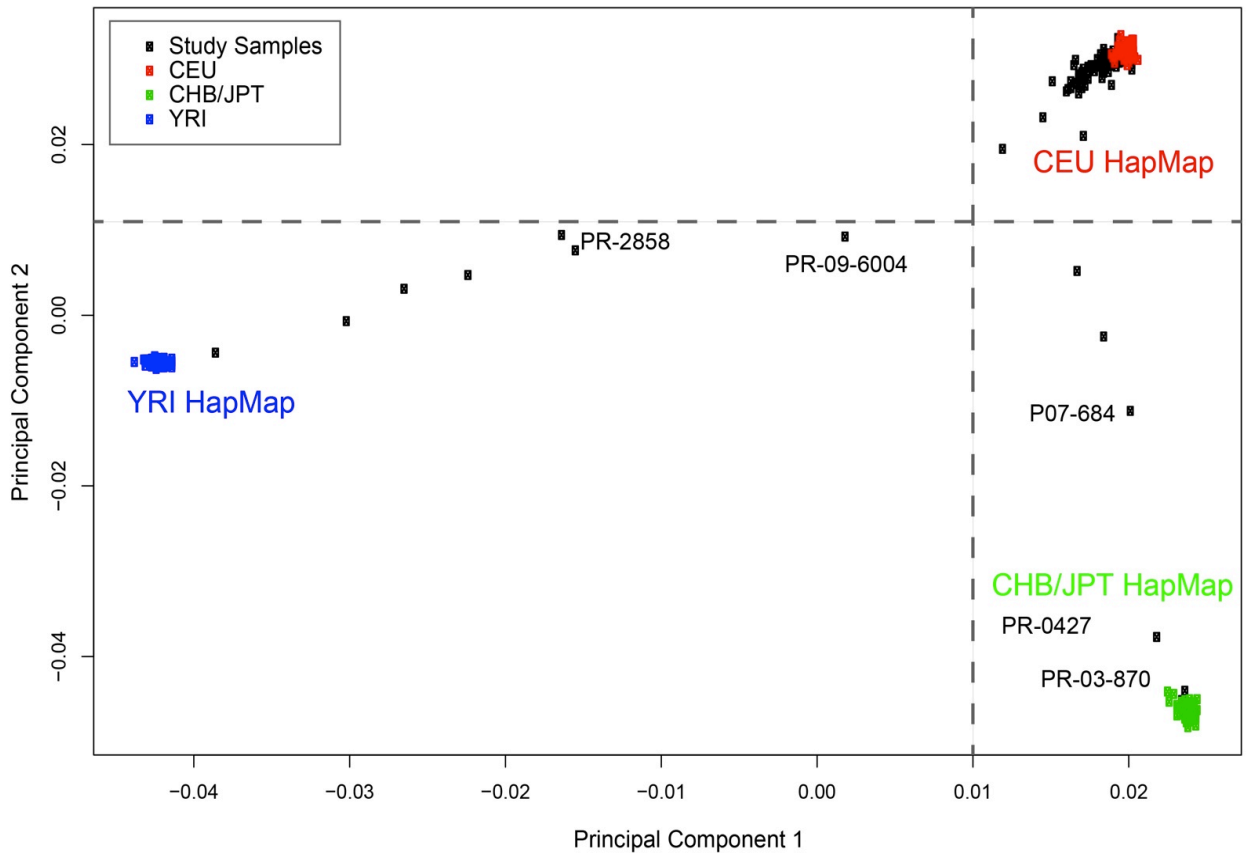


Figure S2.17. Ethnicity analysis of exome-sequenced DNA

Principal component analysis was performed to assess the origin of the study individuals using SNP array data. HapMap Phase II samples representing three distinct populations, European (CEU) (red), Yoruban (YRI) (blue) and Chinese/Japanese (CHB/JPT) (green), were included in the analysis. The study identifiers of the exome-sequenced individuals whose genetic profiles deviate from the CEU pattern are shown.

Figure S3.1. Recurrent rearrangements alter known and putative prostate cancer genes

(**A**) Schematic of *CRKL-MAPK1* and *NRF1-BRAF* fusions detected by WGS, along with validation by FISH assay.

(**B**) Protein-Protein Interaction (PPI) data were analyzed to nominate rearrangements of potential biological consequence. The centrality in a PPI network (Szklarczyk et al., 2011) was assessed for protein products of genes that were rearranged in more than one sample (total 397). X- and Y-axes measure two indexes of centrality, where larger values indicate more central network positions. Circle color and size are proportional to the frequency of gene rearrangement across the tumor cohort. Genes scoring in the 95th percentile are depicted as filled circles. The two panels on the right show the centrality of recurrently rearranged genes (depicted as red circles in the bottom plot) compared to the entire PPI dataset.

(**C**) Disruptive genomic rearrangement of *JAK1*, *JAK2* and *GSK3B*. Dotted lines show intragenic breakpoints and corresponding text indicates the locus to which the breakpoint is fused (IGR; inter-genic region). Rearrangements depicted above the gene diagrams occurred in a sense-preserving orientation; rearrangements below gene diagrams occurred in an anti-sense orientation. Right, genomic rearrangements were validated by FISH.

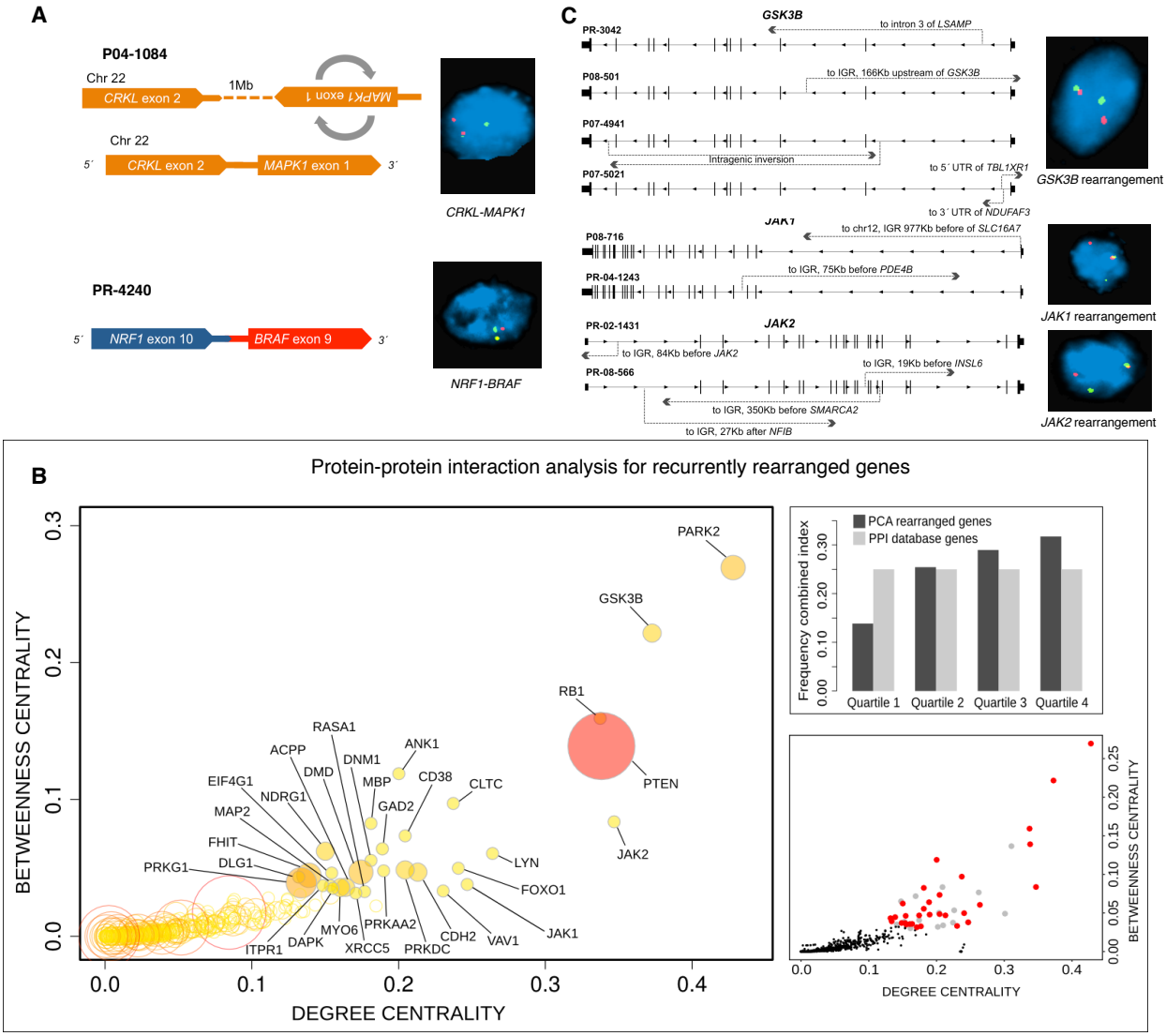


Figure S3.1 (continued)

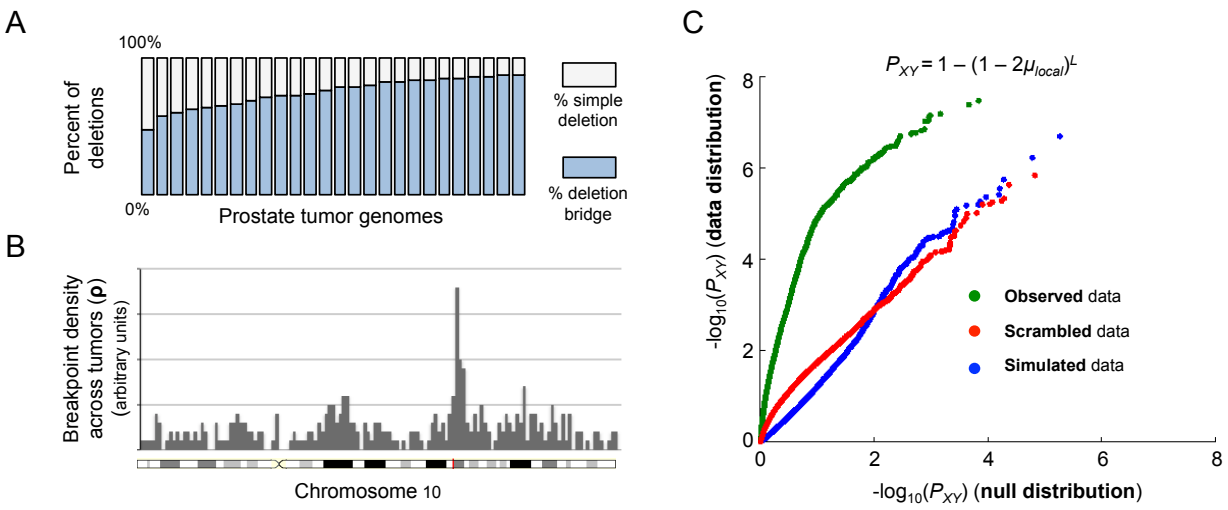


Figure S3.2. Signatures of coordinately generated rearrangement chains

(**A**) Percentage of DNA deletions bounded by fusion breakpoints that were uniquely identified as deletion bridges (blue) or simple deletions (white) in prostate tumors with ten or more deletions in either category.

(**B**) Probabilistic model of independent rearrangements across the genome. The expected distribution of independent DNA breaks in a given tumor (ρ) is estimated by counting the number of tumors with one or more rearrangements within 1Mb tiling windows across the genome. ρ is used to calculate the value of μ_{local} used by ChainFinder in the null model of independent breakpoints.

(**C**) Quantile-quantile (Q-Q) plot comparing the distribution of P_{XY} values (the adjacency probabilities for independent breakpoints) for observed, simulated and scrambled rearrangements.

Figure S3.3. ChainFinder analysis of cancer genomes

(**A**) ChainFinder creates a graph representation of somatic rearrangement breakpoints and corresponding deletions (see Methods for an extended explanation). (1) Breakpoints of somatic fusions are represented as nodes connected by edges. (2) The adjacency probability (P_{XY}) is calculated for pairs of neighboring breakpoints based on their reference genome distance (L) and the local rate of rearrangements (μ_{local}). (3) Breakpoints at either boundary of a deletion bridge are joined by edges. (4) The graph is searched for cycles connecting breakpoints that are unlikely to have arisen independently, based on P_{XY} values of corresponding intervals. (5) The final graph contains sets of rearrangements and associated deletions that are unlikely to have occurred independently.

(**B**) For a hypothetical cycle involving three rearrangements, the independent breakpoint model constitutes all scenarios by which any rearrangement could have arisen independently of others in the cycle (see Methods).

(**C**-**D**) Circos plots of rearrangements color-coded by chain for 57 prostate tumors and 59 previously sequenced cancer genomes (see Table S3.5B for references). Rearrangements in gray were not assigned to a chain. Copy number alteration is shown in blue (deletion) and red (amplification) in the inner ring of each plot.

Figure S3.3, continued

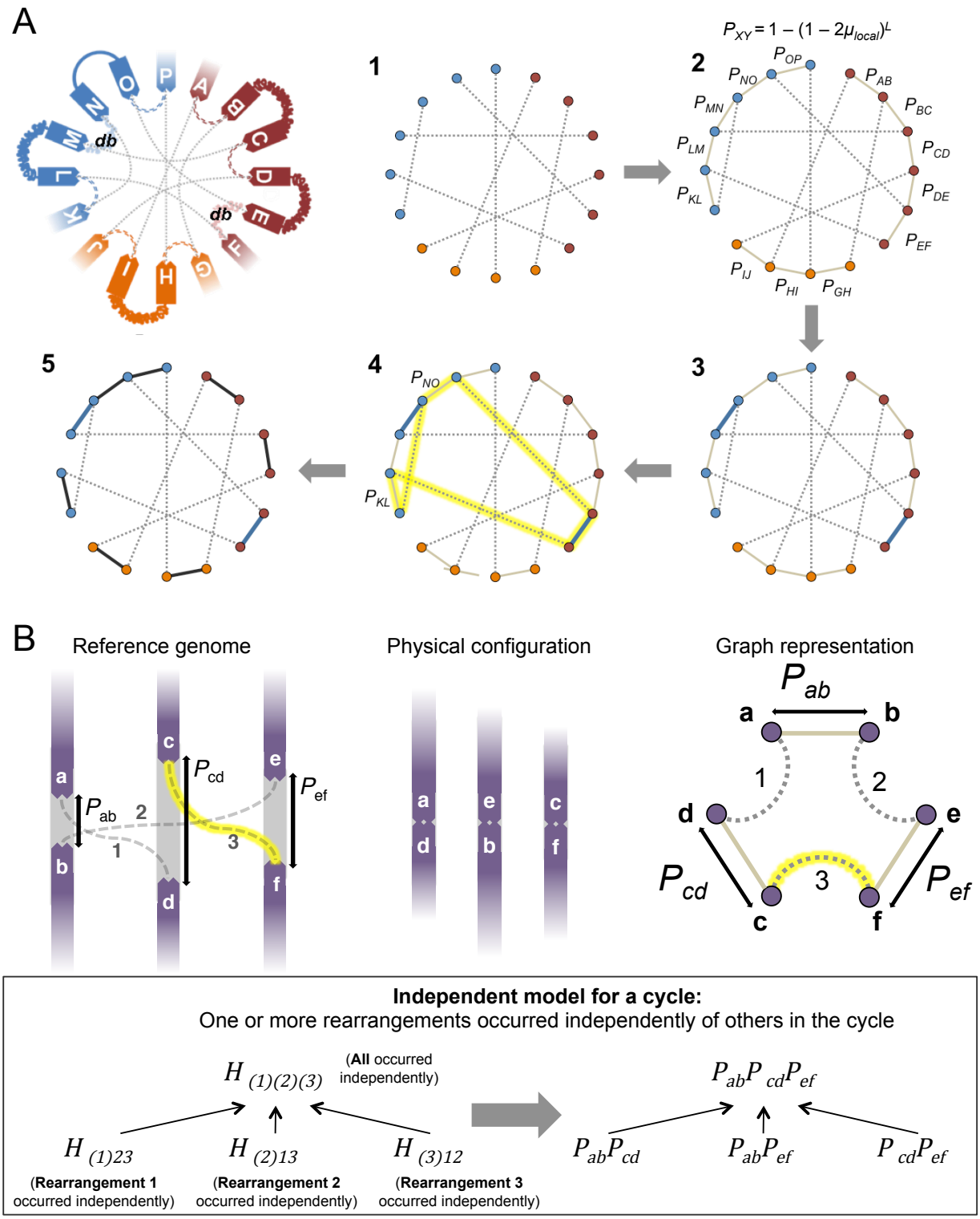
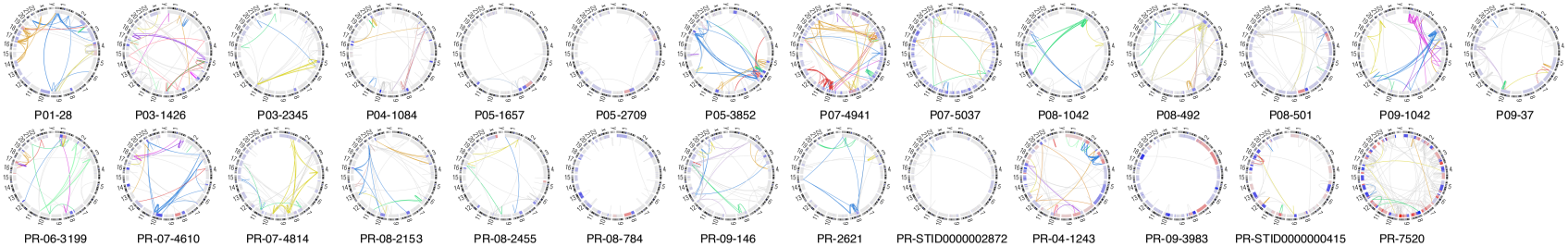


Figure S3.3, continued

C

Prostate cancer (ETS+)



Prostate cancer (ETS-)

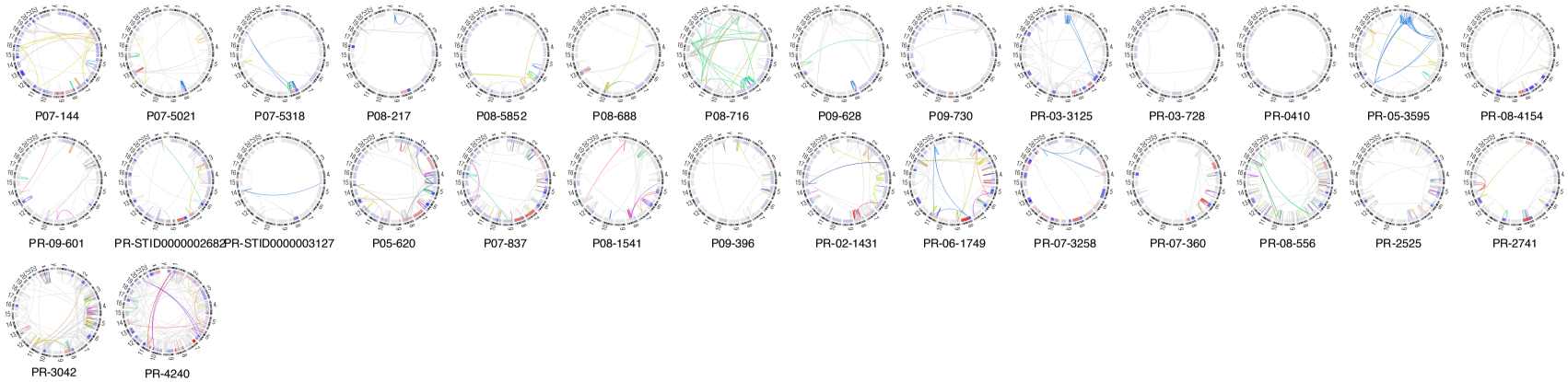


Figure S3.3, continued

D

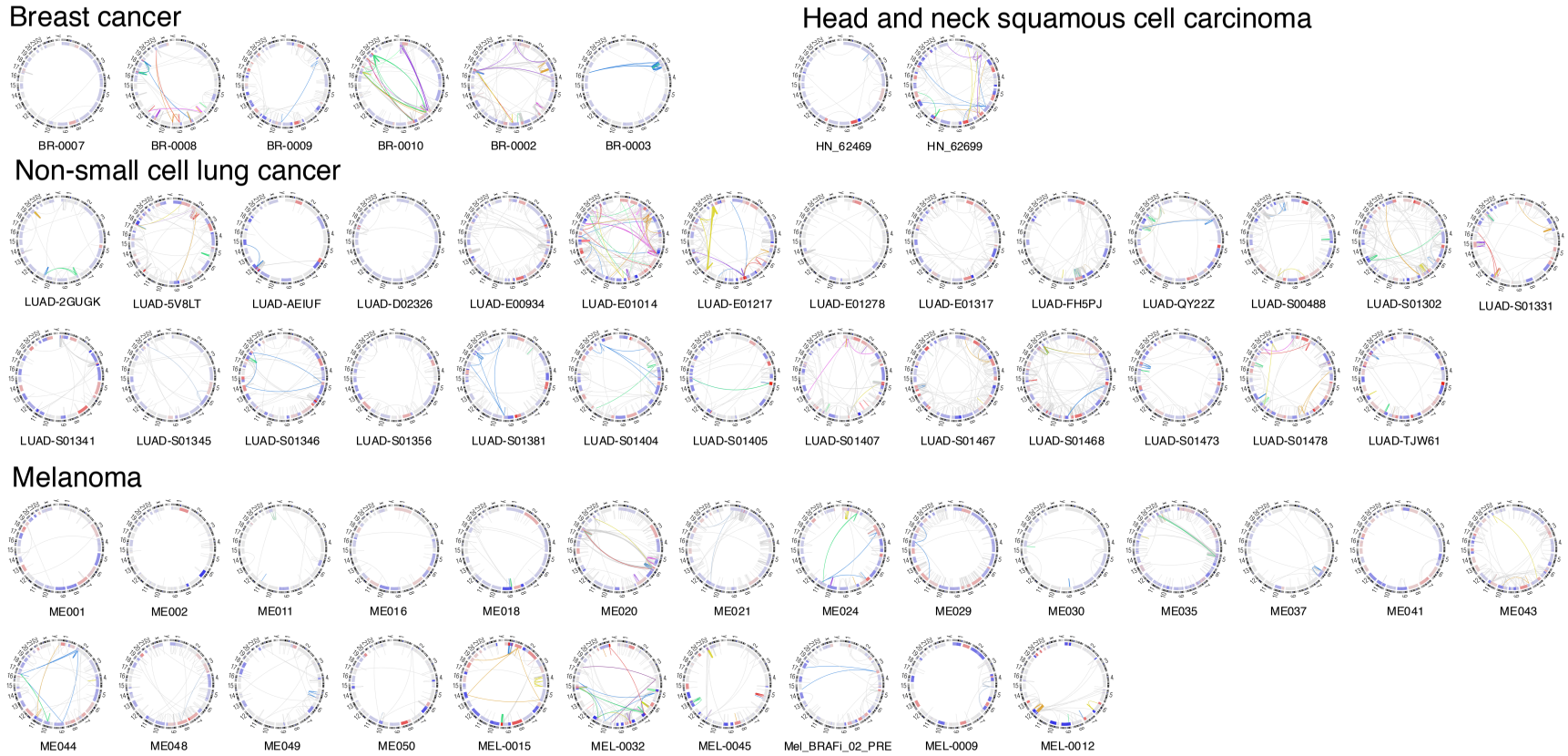


Figure S3.4. Rearrangement profiles of prostate tumor genomes

(**A**) Chromoplexy arises in physically interacting chromatin. Chains are enriched for rearrangements that fuse portions of the genome in close physical proximity as determined by Hi-C analysis of the RWPE-1 prostate epithelial cell line (Rickman et al., 2012). See Methods for further details.

(**B**) Enrichment of rearrangement breakpoints near to and distant from various genomic features, including ChIP-seq peaks from ERG^{+} VCaP prostate cancer cells (Yu et al., 2010). Color hue reflects the degree of enrichment (red) or depletion (blue) and box area reflects statistical significance. “Near” and “Far” correspond to within 100kb and further than 1Mb, respectively. The number of rearrangements for each tumor is depicted in the gray columns (see Methods).

(**C**) Recurrent somatic copy number alterations (SCNAs) across an extended panel of 199 prostate tumors grouped by *CHD1* deletion status. For comparison, the same samples are also grouped by *TP53* deletion status.

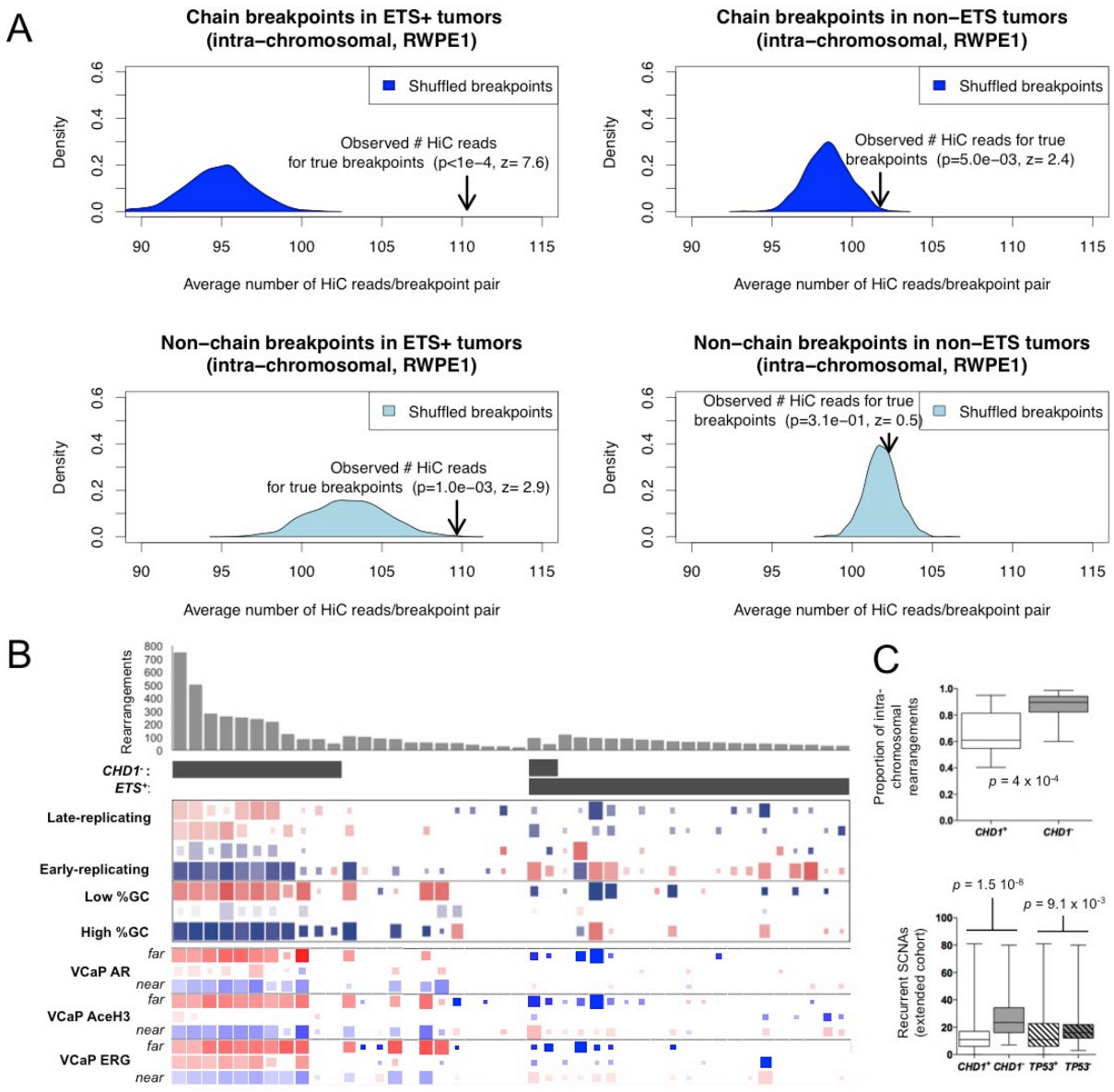


Figure S3.4 (continued)

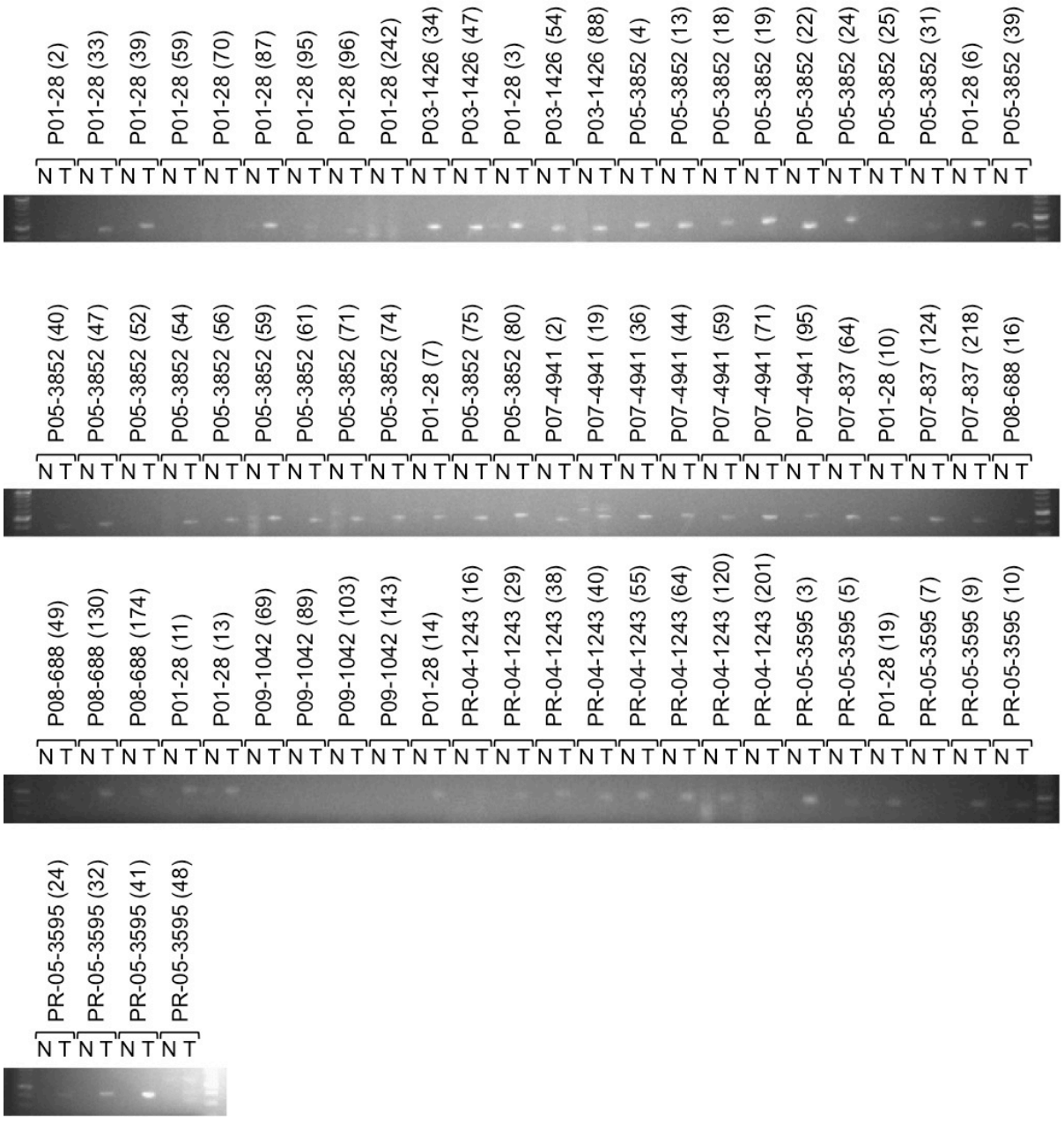


Figure S3.5. PCR validation of chained DNA rearrangements

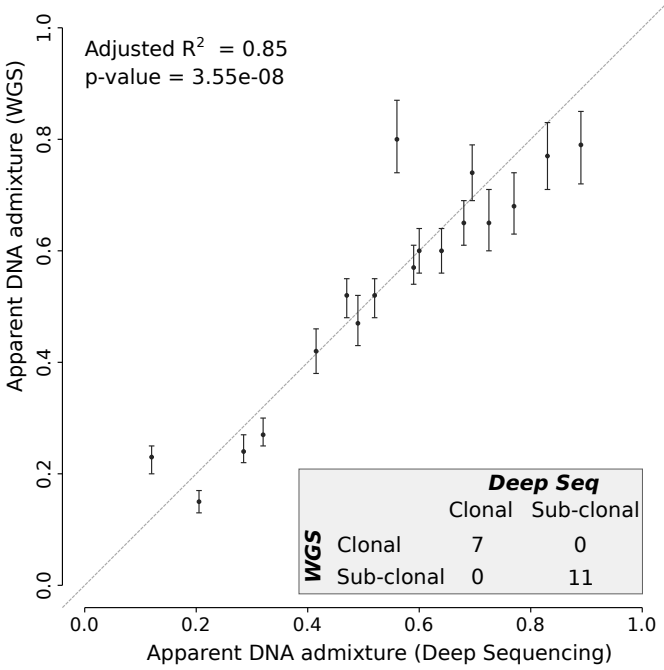
PCR reactions were run on tumor and normal DNA to amplify across the junctions of 76 somatic fusions. Rearrangements are numbered as in Table S3C. Please see Table S3C for a list of additional rearrangements that were validated by PCR and deep sequencing on the MiSeq platform.

Figure S3.6. Estimation of clonality and stromal DNA admixture

(**A**) Apparent DNA admixture computed from WGS and MiSeq deep-sequencing data for 18 somatically deleted genes in 7 samples. Error bars for WGS estimations are computed according to Table S6A. Clonality calls on WGS data were made with a minimum of 20 informative hemizygous SNPs covered to an average depth of 20x or greater. MiSeq calls are based on 4 SNPs with average local coverage of >65,000x. The contingency table (bottom-right) shows the agreement for clonality and sub-clonality calls between MiSeq and WGS based data (Cochran test, p-value = 1).

(**B**) Clonal status of deletions at 14 loci inferred across 49 prostate cancers. The central panel denotes the clonal status of a gene lesion in a sample. Empty dark gray rectangles indicate either that the gene was not deleted or that there were insufficient informative SNPs to determine clonality status. White circles indicate a 100% clonal deletion. Colored circles indicate sub-clonal deletions, where darker color indicates a more subclonal deletion. Top rows report Gleason scores, ranging from 6 (light blue) to 9 (violet), and global stromal DNA admixture, where darker color signifies more admixture. Green bars summarize lesion clonality on a per-sample and per-gene basis. Dark and light green denote the proportion of clonal and sub-clonal deletions, respectively.

A.



B.

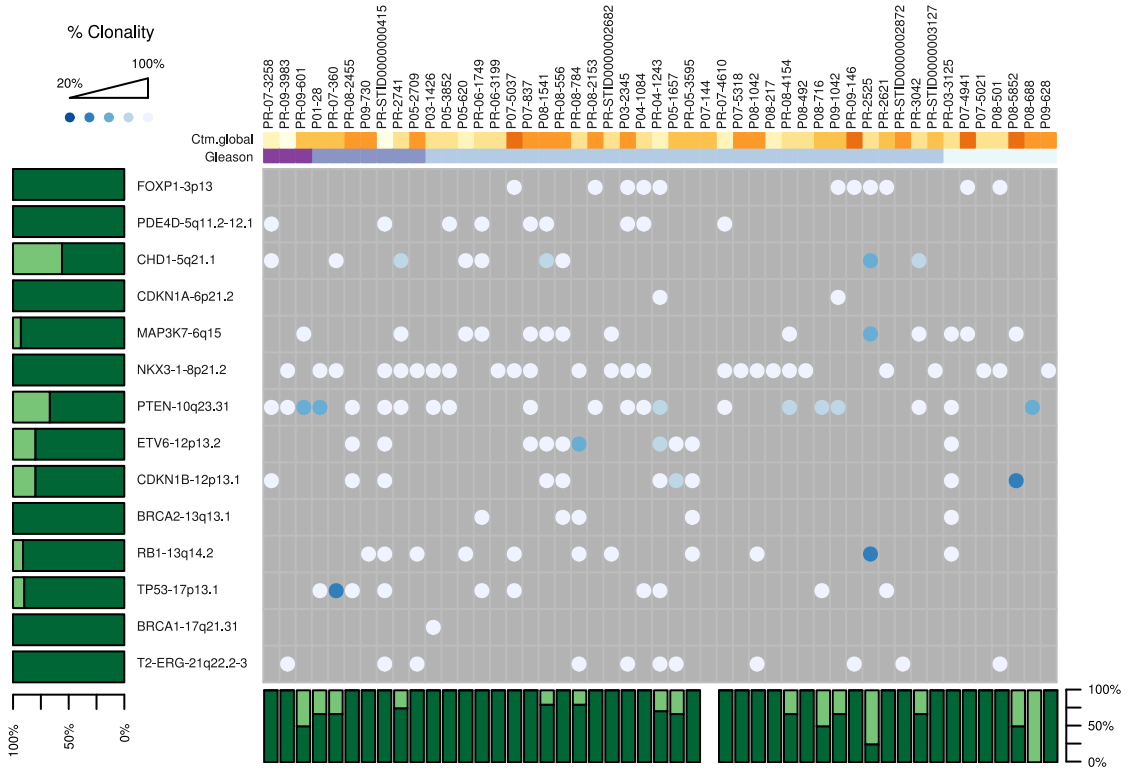


Figure S3.6, continued

Figure S3.7. Chromoplexy continues during outgrowth of tumor sub-clones and may generate multiple rearrangements in closed chains at once

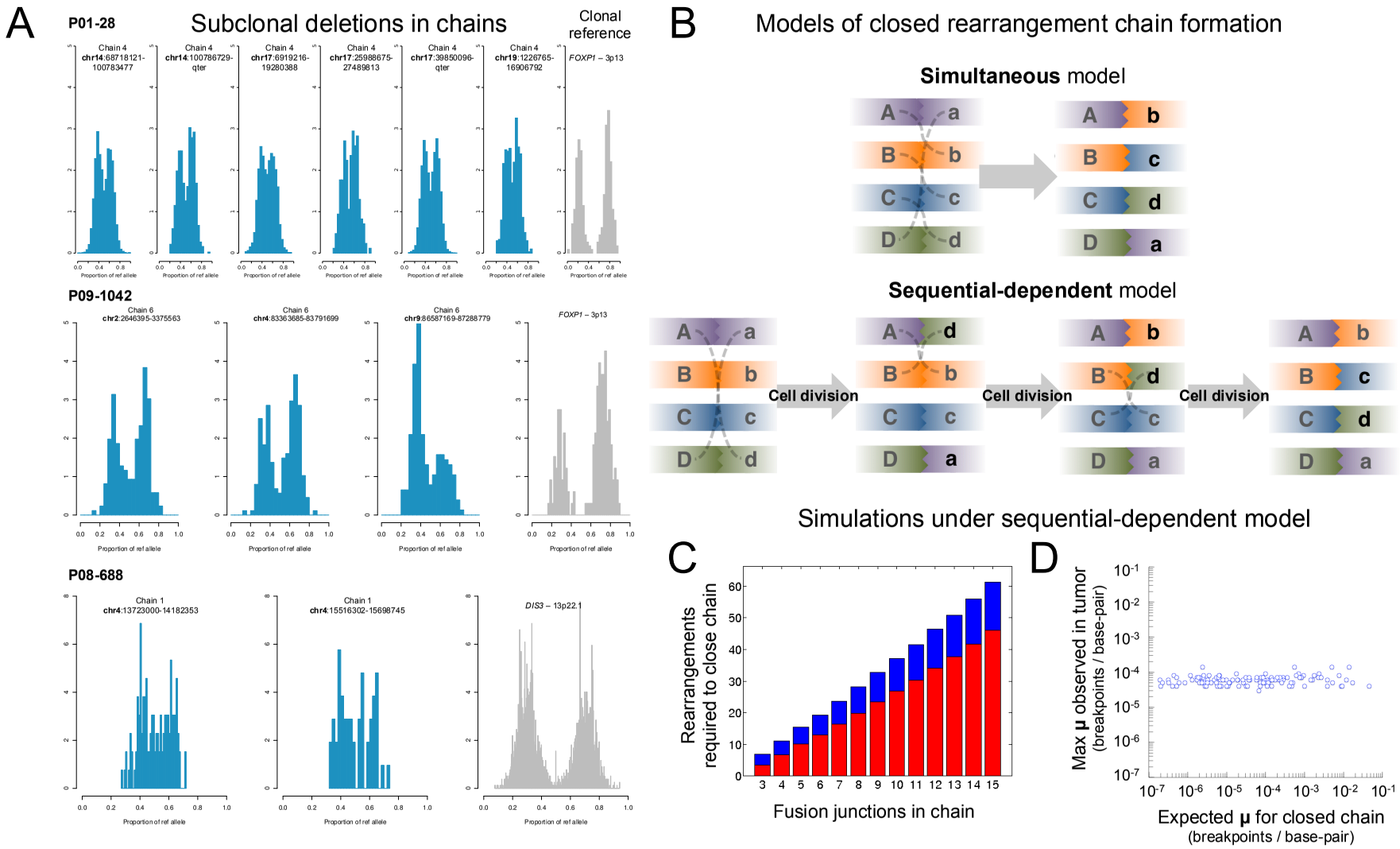
(**A**) Three examples of subclonal chains identified by clonality analysis of deletion bridges. Allelic fraction distributions of heterozygous SNPs within the deleted segments are indicated. On the right, a clonal deletion bridge from the same sample is shown for comparison.

(**B**) Closed chains of non-independent rearrangements could arise from (1) a series of balanced translocations over multiple cell generations (the “sequential-dependent model”) or (2) concerted rearrangements within one cell cycle (the “simultaneous model”).

(**C**) For closed chains of rearrangements, bars indicate the median number of sequential balanced translocations required to close a chain under the sequential-dependent model (assuming translocations occur randomly between breakpoints within the chain). Average values from 10,000 simulations per chain size are shown in blue. Red bars indicate the number of rearrangements that disrupt a previously formed fusion junction.

(**D**) For 121 observed closed chains, the values from (C) and genomic distances between chain breakpoints were used to calculate the local rate of rearrangements required to close the chain under the sequential-dependent model. This density is compared to the maximum density of rearrangements observed in the tumor containing the chain (assessed in 10kb windows genome-wide).

Figure S3.7 (continued)



References

Acloque, H., Adams, M.S., Fishwick, K., Bronner-Fraser, M., and Nieto, M.A. (2009). Epithelial-mesenchymal transitions: the importance of changing cell state in development and disease. J Clin Invest *119*, 1438-1449.

Agell, L., Hernandez, S., Salido, M., de Muga, S., Juanpere, N., Arumi-Uria, M., Menendez, S., Lorenzo, M., Lorente, J.A., Serrano, S.*, et al.* (2011). PI3K signaling pathway is activated by PIK3CA mRNA overexpression and copy gain in prostate tumors, but PIK3CA, BRAF, KRAS and AKT1 mutations are infrequent events. Mod Pathol *24*, 443-452.

Ahmad, I., Patel, R., Singh, L.B., Nixon, C., Seywright, M., Barnetson, R.J., Brunton, V.G., Muller, W.J., Edwards, J., Sansom, O.J.*, et al.* (2011). HER2 overcomes PTEN (loss)-induced senescence to cause aggressive prostate cancer. Proc Natl Acad Sci U S A *108*, 16392-16397.

Alers, J.C., Rochat, J., Krijtenburg, P.J., Hop, W.C., Kranse, R., Rosenberg, C., Tanke, H.J., Schroder, F.H., and van Dekken, H. (2000). Identification of genetic markers for prostatic cancer progression. Lab Invest *80*, 931-942.

Andreoiu, M., and Cheng, L. (2010). Multifocal prostate cancer: biologic, prognostic, and therapeutic implications. Hum Pathol *41*, 781-793.

Asatiani, E., Huang, W.X., Wang, A., Rodriguez Ortner, E., Cavalli, L.R., Haddad, B.R., and Gelmann, E.P. (2005). Deletion, methylation, and expression of the NKX3.1 suppressor gene in primary human prostate cancer. Cancer Res *65*, 1164-1173.

Baca, S.C., and Garraway, L.A. (2012). The genomic landscape of prostate cancer. Front Endocrinol (Lausanne) *3*, 69.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M, Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi*,* M., *et al.* (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666-77.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N.*, et al.* (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet *44*, 685-689.

Benjamini, Y.a.H., Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc *57*, 289-300.

Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C.*, et al.* (2011). The genomic complexity of primary human prostate cancer. Nature *470*, 214-220.

Berger, R., Febbo, P.G., Majumder, P.K., Zhao, J.J., Mukherjee, S., Signoretti, S., Campbell, K.T., Sellers, W.R., Roberts, T.M., Loda, M.*, et al.* (2004). Androgen-induced differentiation and tumorigenicity of human prostate epithelial cells. Cancer Res *64*, 8867-8875.

Bergethon, K., Shaw, A.T., Ou, S.H., Katayama, R., Lovly, C.M., McDonald, N.T., Massion, P.P., Siwak-Tapp, C., Gonzalez, A., Fang, R.*, et al.* (2012). ROS1 rearrangements define a unique molecular class of lung cancers. J Clin Oncol *30*, 863-870.

Beroud, C., and Soussi, T. (1996). APC gene: database of germline and somatic mutations in human tumors and cell lines. Nucleic Acids Res *24*, 121-124.

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S.*, et al.* (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A *104*, 20007-20012.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M.*, et al.* (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899-905.

Bhatia-Gaur, R., Donjacour, A.A., Sciavolino, P.J., Kim, M., Desai, N., Young, P., Norton, C.R., Gridley, T., Cardiff, R.D., Cunha, G.R.*, et al.* (1999). Roles for Nkx3.1 in prostate development and cancer. Genes Dev *13*, 966-977.

Bookstein, R., MacGrogan, D., Hilsenbeck, S.G., Sharkey, F., and Allred, D.C. (1993). p53 is mutated in a subset of advanced-stage prostate cancers. Cancer Res *53*, 3369-3373.

Boormans, J.L., Korsten, H., Ziel-van der Made, A.C., van Leenders, G.J., Verhagen, P.C., and Trapman, J. (2010). E17K substitution in AKT1 in prostate cancer. Br J Cancer *102*, 1491-1494.

Bowen, C., and Gelmann, E.P. (2010). NKX3.1 activates cellular response to DNA damage. Cancer Res *70*, 3089-3097.

Brenner, J.C., Ateeq, B., Li, Y., Yocum, A.K., Cao, Q., Asangani, I.A., Patel, S., Wang, X., Liang, H., Yu, J.*, et al.* (2011). Mechanistic rationale for inhibition of poly(ADP-ribose) polymerase in ETS gene fusion-positive prostate cancer. Cancer cell *19*, 664-678.

Brewster, S.F., Browne, S., and Brown, K.W. (1994). Somatic allelic loss at the DCC, APC, nm23-H1 and p53 tumor suppressor gene loci in human prostatic carcinoma. J Urol *151*, 1073-1077.

Brothman, A.R., Maxwell, T.M., Cui, J., Deubler, D.A., and Zhu, X.L. (1999). Chromosomal clues to the development of prostate tumors. Prostate *38*, 303-312.

Brown, K.R., and Jurisica, I. (2007). Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol *8*, R95.

Buhrman, G., Wink, G., and Mattos, C. (2007). Transformation efficiency of RasQ61 mutants linked to structural features of the switch regions in the presence of Raf. Structure 15, 1618-1629.

Burrow, A.A., Marullo, A., Holder, L.R., and Wang, Y.H. (2010). Secondary structure formation and DNA instability at fragile site FRA16B. Nucleic Acids Res *38*, 2865-2877.

Cai, C.Q., Peng, Y., Buckley, M.T., Wei, J., Chen, F., Liebes, L., Gerald, W.L., Pincus, M.R., Osman, I., and Lee, P. (2008). Epidermal growth factor receptor activation in prostate cancer by three novel missense mutations. Oncogene *27*, 3201-3210.

Cairns, P., Okami, K., Halachmi, S., Halachmi, N., Esteller, M., Herman, J.G., Jen, J., Isaacs, W.B., Bova, G.S., and Sidransky, D. (1997a). Frequent inactivation of PTEN/MMAC1 in primary prostate cancer. Cancer Res *57*, 4997-5000.

Camidge, D.R., Bang, Y.J., Kwak, E.L., Iafrate, A.J., Varella-Garcia, M., Fox, S.B., Riely, G.J., Solomon, B., Ou, S.H., Kim, D.W.*, et al.* (2012). Activity and safety of crizotinib in patients with ALK-positive non-small-cell lung cancer: updated results from a phase 1 study. Lancet Oncol *13*, 1011-1019.

The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Canman, C.E., Lim, D.S., Cimprich, K.A., Taya, Y., Tamai, K., Sakaguchi, K., Appella, E., Kastan, M.B., and Siliciano, J.D. (1998). Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. Science 281, 1677-1679.

Carpten, J.D., Faber, A.L., Horn, C., Donoho, G.P., Briggs, S.L., Robbins, C.M., Hostetter, G., Boguslawski, S., Moses, T.Y., Savage, S., et al. (2007). A transforming mutation in the pleckstrin homology domain of AKT1 in cancer. Nature 448, 439-444.

Carrano, A.C., Eytan, E., Hershko, A., and Pagano, M. (1999). SKP2 is required for ubiquitin-mediated degradation of the CDK inhibitor p27. Nat Cell Biol *1*, 193-199.

Carter, B.S., Epstein, J.I., and Isaacs, W.B. (1990). Ras gene mutations in human prostate cancer. Cancer Res 50, 6830-6832.

Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A.*, et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol *30*, 413-421.

Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K., Sougnez, C., Schinzel, A.C., Harview, C.L., Brunet, J.P., Ahmann, G.J., Adli, M.*, et al.* (2011). Initial genome sequencing and analysis of multiple myeloma. Nature *471*, 467-472.

Chen, M., Pratt, C.P., Zeeman, M.E., Schultz, N., Taylor, B.S., O'Neill, A., Castillo-Martin, M., Nowak, D.G., Naguib, A., Grace, D.M.*, et al.* (2011). Identification of PHLPP1 as a tumor suppressor reveals the role of feedback activation in PTEN-mutant prostate cancer progression. Cancer Cell *20*, 173-186.

Chen, Y., Wang, J., Fraig, M.M., Metcalf, J., Turner, W.R., Bissada, N.K., Watson, D.K., and Schweinfest, C.W. (2001). Defects of DNA mismatch repair in human prostate cancer. Cancer Res *61*, 4112-4121.

Chen, Z., Trotman, L.C., Shaffer, D., Lin, H.K., Dotan, Z.A., Niki, M., Koutcher, J.A., Scher, H.I., Ludwig, T., Gerald, W.*, et al.* (2005). Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. Nature *436*, 725-730.

Chesire, D.R., Ewing, C.M., Sauvageot, J., Bova, G.S., and Isaacs, W.B. (2000). Detection and analysis of beta-catenin mutations in prostate cancer. Prostate *45*, 323-334.

Cho, N.Y., Choi, M., Kim, B.H., Cho, Y.M., Moon, K.C., and Kang, G.H. (2006). BRAF and KRAS mutations in prostatic adenocarcinoma. Int J Cancer *119*, 1858-1862.

Chou, J., Provot, S., and Werb, Z. (2010). GATA3 in development and cancer differentiation: cells GATA have it! J Cell Physiol *222*, 42-49.

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., and Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. Bioinformatics *27*, 2601-2602.

Clark, K.L., Halay, E.D., Lai, E., and Burley, S.K. (1993). Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature *364*, 412-420.

Cook, P.R. (2010). A model for all genomes: the role of transcription factories. J Mol Biol *395*, 1-10.

Cronauer, M.V., Schulz, W.A., Ackermann, R., and Burchardt, M. (2005). Effects of WNT/beta-catenin pathway activation on signaling through T-cell factor and androgen receptor in prostate cancer cell lines. Int J Oncol *26*, 1033-1040.

Dahiya, R., Lee, C., McCarville, J., Hu, W., Kaur, G., and Deng, G. (1997). High frequency of genetic instability of microsatellites in human prostatic adenocarcinoma. Int J Cancer *72*, 762-767.

Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al. (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 462, 739-744.

Daskivich, T.J., Chamie, K., Kwan, L., Labo, J., Palvolgyi, R., Dash, A., Greenfield, S., and Litwin, M.S. (2011). Overtreatment of men with low-risk prostate cancer and significant comorbidity. Cancer *117*, 2058-2066.

De, S., and Michor, F. (2011). DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat Biotechnol *29*, 1103-1108.

Demichelis, F., Fall, K., Perner, S., Andren, O., Schmidt, F., Setlur, S.R., Hoshida, Y., Mosquera, J.M., Pawitan, Y., Lee, C*., et al.* (2007). TMPRSS2:ERG gene fusion associated with lethal prostate cancer in a watchful waiting cohort. Oncogene *26*, 4596-4599.

Demichelis, F., Setlur, S.R., Beroukhim, R., Perner, S., Korbel, J.O., Lafargue, C.J., Pflueger, D., Pina, C., Hofer, M.D., Sboner, A*., et al.* (2009). Distinct genomic aberrations associated with ERG rearranged prostate cancer. Genes Chromosomes Cancer *48*, 366-380.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M*., et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

Di Cristofano, A., De Acetis, M., Koff, A., Cordon-Cardo, C., and Pandolfi, P.P. (2001). Pten and p27KIP1 cooperate in prostate cancer tumor suppression in the mouse. Nat Genet *27*, 222-224.

Di, L.J., Byun, J.S., Wong, M.M., Wakano, C., Taylor, T., Bilke, S., Baek, S., Hunter, K., Yang, H., Lee, M*., et al.* (2013). Genome-wide profiles of CtBP link metabolism with genome stability and epithelial reprogramming in breast cancer. Nat Commun *4*, 1449.

Ding, Z., Wu, C.J., Chu, G.C., Xiao, Y., Ho, D., Zhang, J., Perry, S.R., Labrot, E.S., Wu, X., Lis, R*., et al.* (2011). SMAD4-dependent barrier constrains prostate cancer growth and metastatic progression. Nature *470*, 269-273.

Dong, J.T. (2006). Prevalent mutations in prostate cancer. J Cell Biochem *97*, 433-447.

Dong, X.Y., Chen, C., Sun, X., Guo, P., Vessella, R.L., Wang, R.X., Chung, L.W., Zhou, W., and Dong, J.T. (2006). FOXO1A is a candidate for the 13q14 tumor suppressor gene inhibiting androgen receptor signaling in prostate cancer. Cancer Res *66*, 6998-7006.

Donner, A.J., Szostek, S., Hoover, J.M., and Espinosa, J.M. (2007). CDK8 is a stimulus-specific positive coregulator of p53 target genes. Mol Cell *27*, 121-133.

Dreher, T., Zentgraf, H., Abel, U., Kappeler, A., Michel, M.S., Bleyl, U., and Grobholz, R. (2004). Reduction of PTEN and p27kip1 expression correlates with tumor grade in prostate cancer. Analysis in radical prostatectomy specimens and needle biopsies. Virchows Arch *444*, 509-517.

Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., Beroukhim, R., and Getz, G. (2012). Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res *2*, 228-35.

Eastham, J.A., Stapleton, A.M., Gousse, A.E., Timme, T.L., Yang, G., Slawin, K.M., Wheeler, T.M., Scardino, P.T., and Thompson, T.C. (1995). Association of p53 mutations with metastatic prostate cancer. Clin Cancer Res *1*, 1111-1118.

Edwards, J., Krishna, N.S., Witton, C.J., and Bartlett, J.M. (2003). Gene amplifications associated with the development of hormone-resistant prostate cancer. Clin Cancer Res *9*, 5271-5281.

Elliott, R.L., and Blobe, G.C. (2005). Role of transforming growth factor Beta in human cancer. J Clin Oncol *23*, 2078-2093.

Emmert-Buck, M.R., Vocke, C.D., Pozzatti, R.O., Duray, P.H., Jennings, S.B., Florence, C.D., Zhuang, Z., Bostwick, D.G., Liotta, L.A., and Linehan, W.M. (1995). Allelic loss on chromosome 8p12-21 in microdissected prostatic intraepithelial neoplasia. Cancer Res *55*, 2959-2962.

Espina, V., Wulfkuhle, J.D., Calvert, V.S., VanMeter, A., Zhou, W., Coukos, G., Geho, D.H., Petricoin, E.F., 3rd, and Liotta, L.A. (2006). Laser-capture microdissection. Nat Protoc *1*, 586-603.

Fang, M., Li, J., Blauwkamp, T., Bhambhani, C., Campbell, N., and Cadigan, K.M. (2006). C-terminal-binding protein directly activates and represses Wnt transcriptional targets in Drosophila. Embo J *25*, 2735-2745.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759-767.

Febbo, P.G. (2009). Genomic approaches to outcome prediction in prostate cancer. Cancer *115*, 3046-3057.

Ferrero, D., Tarella, C., Gallo, E., Ruscetti, F.W., and Breitman, T.R. (1982). Terminal differentiation of the human promyelocytic leukemia cell line, HL-60, in the absence of cell proliferation. Cancer Res *42*, 4421-4426.

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L.*, et al.* (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol *12*, R1.

Forbes, S.A., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A.*, et al.* (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res *39*, D945-950.

Forment, J.V., Kaidi, A., and Jackson, S.P. (2012). Chromothripsis and cancer: causes and consequences of chromosome shattering. Nat Rev Cancer *12*, 663-670.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res *39*, D876-882.

Gaddipati, J.P., McLeod, D.G., Heidenberg, H.B., Gaddipatl, J.P., Sesterhenn, I., Finger, M., Moul, J.W., and Srivastava, S. (1994). Frequent detection of codon 877 mutation in the androgen receptor gene in advanced prostate cancers. Cancer Res, 2861-2864.

Gao, H., Ouyang, X., Banach-Petrosky, W.A., Gerald, W.L., Shen, M.M., and Abate-Shen, C. (2006). Combinatorial activities of Akt and B-Raf/Erk signaling in a mouse model of androgen-independent prostate cancer. Proc Natl Acad Sci U S A *103*, 14477-14482.

Gao, L., and Alumkal, J. (2010). Epigenetic regulation of androgen receptor signaling in prostate cancer. Epigenetics *5*, 100-104.

Gao, N., Ishii, K., Mirosevich, J., Kuwajima, S., Oppenheimer, S.R., Roberts, R.L., Jiang, M., Yu, X., Shappell, S.B., Caprioli, R.M.*, et al.* (2005). Forkhead box A1 regulates prostate ductal morphogenesis and promotes epithelial cell maturation. Development *132*, 3431-3443.

Gao, N., Zhang, J., Rao, M.a., Case, T.C., Mirosevich, J., Wang, Y., Jin, R., Gupta, A., Rennie, P.S., and Matusik, R.J. (2003). The role of hepatocyte nuclear factor-3 alpha (Forkhead Box A1) and androgen receptor in transcriptional regulation of prostatic genes. Mol Endocrinol *17*, 1484-1507.

Gaspar-Maia, A., Alajem, A., Polesso, F., Sridharan, R., Mason, M.J., Heidersbach, A., Ramalho-Santos, J., McManus, M.T., Plath, K., Meshorer, E.*, et al.* (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. Nature *460*, 863-868.

Getz, G., Hofling, H., Mesirov, J.P., Golub, T.R., Meyerson, M., Tibshirani, R., and Lander, E.S. (2007). Comment on "The consensus coding sequences of human breast and colorectal cancers". Science *317*, 1500.

Gioeli, D., Mandell, J.W., Petroni, G.R., Frierson, H.F., Jr., and Weber, M.J. (1999). Activation of mitogen-activated protein kinase associated with prostate cancer progression. Cancer Res *59*, 279-284.

Glinsky, G.V., Glinskii, A.B., Stephenson, A.J., Hoffman, R.M., and Gerald, W.L. (2004). Gene expression profiling predicts clinical outcome of prostate cancer. J Clin Invest *113*, 913-923.

Gould, S.J., Eldrige, Niles (1977). Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology *3*, 115-151.

Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C.*, et al.* (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature *487*, 239-243.

Gray, I.C., Stewart, L.M., Phillips, S.M., Hamilton, J.A., Gray, N.E., Watson, G.J., Spurr, N.K., and Snary, D. (1998). Mutation and expression analysis of the putative prostate tumour-suppressor gene PTEN. Br J Cancer *78*, 1296-1300.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C.*, et al.* (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153-158.

Griffith, O.L., Montgomery, S.B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M.C., Bilenky, M., Haeussler, M.*, et al.* (2008). ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res *36*, D107-113.

Guan, M., Xu, C., Zhang, F., and Ye, C. (2009). Aberrant methylation of EphA7 in human prostate cancer and its relation to clinicopathologic features. Int J Cancer 124, 88-94.

Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L., et al. (2011). Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat Genet 43, 875-878.

Habegger, L., Sboner, A., Gianoulis, T.A., Rozowsky, J., Agarwal, A., Snyder, M., and Gerstein, M. (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics *27*, 281-283.

Haffner, M.C., Aryee, M.J., Toubaji, A., Esopi, D.M., Albadine, R., Gurel, B., Isaacs, W.B., Bova, G.S., Liu, W., Xu, J.*, et al.* (2010). Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nat Genet *42*, 668-675.

Halvorsen, O.J., Haukaas, S.A., and Akslen, L.A. (2003). Combined loss of PTEN and p27 expression is associated with tumor cell proliferation by Ki-67 and increased risk of recurrent disease in localized prostate cancer. Clinical Cancer Res *9*, 1474-1479.

Harbour, J.W., Onken, M.D., Roberson, E.D., Duan, S., Cao, L., Worley, L.A., Council, M.L., Matatall, K.A., Helms, C., and Bowcock, A.M. (2010). Frequent mutation of BAP1 in metastasizing uveal melanomas. Science *330*, 1410-1413.

He, W.W., Sciavolino, P.J., Wing, J., Augustus, M., Hudson, P., Meissner, P.S., Curtis, R.T., Shell, B.K., Bostwick, D.G., Tindall, D.J.*, et al.* (1997). A novel human prostate-specific, androgen-regulated homeobox gene (NKX3.1) that maps to 8p21, a region frequently deleted in prostate cancer. Genomics *43*, 69-77.

Heidenberg, H.B., Sesterhenn, I.A., Gaddipati, J.P., Weghorst, C.M., Buzard, G.S., Moul, J.W., and Srivastava, S. (1995). Alteration of the tumor suppressor gene p53 in a high fraction of hormone refractory prostate cancer. J Urol *154*, 414-421.

Hernandez-Munoz, I., Lund, A.H., van der Stoop, P., Boutsma, E., Muijrers, I., Verhoeven, E., Nusinow, D.A., Panning, B., Marahrens, Y., and van Lohuizen, M. (2005). Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. Proc Natl Acad Sci USA *102*, 7635-7640.

Holcomb, I.N., Young, J.M., Coleman, I.M., Salari, K., Grove, D.I., Hsu, L., True, L.D., Roudier, M.P., Morrissey, C.M., Higano, C.S.*, et al.* (2009). Comparative analyses of chromosome alterations in soft-tissue metastases within and across patients with castration-resistant prostate cancer. Cancer Res *69*, 7793-7802.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand J Stat *6*, 65-70.

Huang, S., Gulzar, Z.G., Salari, K., Lapointe, J., Brooks, J.D., and Pollack, J.R. (2011). Recurrent deletion of CHD1 in prostate cancer with relevance to cell invasiveness. Oncogene.

Hyytinen, E.R., Frierson, H.F., Jr., Boyd, J.C., Chung, L.W., and Dong, J.T. (1999). Three distinct regions of allelic loss at 13q14, 13q21-22, and 13q33 in prostate cancer. Genes Chromosomes Cancer *25*, 108-114.

Janne, P.A., and Meyerson, M. (2012). ROS1 rearrangements in lung cancer: a new genomic subset of lung adenocarcinoma. J Clin Oncol *30*, 878-879.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. CA Cancer J Clin *61*, 69-90.

Jenkins, R.B., Qian, J., Lieber, M.M., and Bostwick, D.G. (1997). Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization. Cancer Res *57*, 524-531.

Kan, Z., Jaiswal, B.S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H.M., Yue, P., Haverty, P.M., Bourgon, R., Zheng, J.*, et al.* (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. Nature *466*, 869-873.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res *40*, D109-114.

Kennedy, J.E., R. (1995). Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks *IV*, 1942-1948.

Kibel, A.S., Suarez, B.K., Belani, J., Oh, J., Webster, R., Brophy-Ebbers, M., Guo, C., Catalona, W.J., Picus, J., and Goodfellow, P.J. (2003). CDKN1A and CDKN1B polymorphisms and risk of advanced prostate carcinoma. Cancer Res *63*, 2033-2036.

Kim, J., Roh, M., Doubinskaia, I., Algarroba, G.N., Eltoum, I.E., and Abdulkadir, S.A. (2011a). A mouse model of heterogeneous, c-MYC-initiated prostate cancer with loss of Pten and p53. Oncogene.

Kim, J.H., Dhanasekaran, S.M., Prensner, J.R., Cao, X., Robinson, D., Kalyana-Sundaram, S., Huang, C., Shankar, S., Jing, X., Iyer, M.*, et al.* (2011b). Deep sequencing reveals distinct patterns of DNA methylation in prostate cancer. Genome Res *21*, 1028-1041.

Kim, M.J., Cardiff, R.D., Desai, N., Banach-Petrosky, W.A., Parsons, R., Shen, M.M., and Abate-Shen, C. (2002). Cooperativity of Nkx3.1 and Pten loss of function in a mouse model of prostate carcinogenesis. Proc Natl Acad Sci U S A *99*, 2884-2889.

King, J.C., Xu, J., Wongvipat, J., Hieronymus, H., Carver, B.S., Leung, D.H., Taylor, B.S., Sander, C., Cardiff, R.D., Couto, S.S.*, et al.* (2009). Cooperativity of TMPRSS2-ERG with PI3-kinase pathway activation in prostate oncogenesis. Nat Genet *41*, 524-526.

Kinkade, C.W., Castillo-Martin, M., Puzio-Kuter, A., Yan, J., Foster, T.H., Gao, H., Sun, Y., Ouyang, X., Gerald, W.L., Cordon-Cardo, C.*, et al.* (2008). Targeting AKT/mTOR and ERK MAPK signaling inhibits hormone-refractory prostate cancer in a preclinical mouse model. J Clin Invest *118*, 3051-3064.

Koivisto, P., Kononen, J., Palmberg, C., and Cleutjens, K. (1997). Androgen receptor gene amplification: a possible molecular mechanism for androgen deprivation therapy failure in prostate cancer. Cancer Res, 314-319.

Konishi, N., Hiasa, Y., Tsuzuki, T., Tao, M., Enomoto, T., and Miller, G.J. (1997). Comparison of ras activation in prostate carcinoma in Japanese and American men. Prostate *30*, 53-57.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-157.

Kumar, a., White, T.a., MacKenzie, a.P., Clegg, N., Lee, C., Dumpit, R.F., Coleman, I., Ng, S.B., Salipante, S.J., Rieder, M.J.*, et al.* (2011). Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. Proc Natl Acad Sci USA *108*.

Kumar-Sinha, C., Tomlins, S.A., and Chinnaiyan, A.M. (2008). Recurrent gene fusions in prostate cancer. Nat Rev Cancer *8*, 497-511.

Kwabi-Addo, B., Wang, J., Erdem, H., Vaid, A., Castro, P., Ayala, G., and Ittmann, M. (2004). The expression of Sprouty1, an inhibitor of fibroblast growth factor signal transduction, is decreased in human prostate cancer. Cancer Res *64*, 4728-4735.

Kwon, J.E., La, M., Oh, K.H., Oh, Y.M., Kim, G.R., Seol, J.H., Baek, S.H., Chiba, T., Tanaka, K., Bang, O.S.*, et al.* (2006). BTB domain-containing speckle-type POZ protein (SPOP) serves as an adaptor of Daxx for ubiquitination by Cul3-based ubiquitin ligase. J Biol Chem *281*, 12664-12672.

Landau, D.A., Carter, S.L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M.S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L.*, et al.* (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell *152*, 714-726.

Lapointe, J., Li, C., Giacomini, C.P., Salari, K., Huang, S., Wang, P., Ferrari, M., Hernandez-Boussard, T., Brooks, J.D., and Pollack, J.R. (2007). Genomic profiling reveals alternative genetic pathways of prostate tumorigenesis. Cancer Res *67*, 8504-8510.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U.*, et al.* (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci U S A *101*, 811-816.

Li, C., Ao, J., Fu, J., Lee, D.F., Xu, J., Lonard, D., and O'Malley, B.W. (2011). Tumor-suppressor role for the SPOP ubiquitin ligase in signal-dependent proteolysis of the oncogenic co-activator SRC-3/AIB1. Oncogene *30*, 4350-4364.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, J. (1997). PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science *275*, 1943-1947.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.G., Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K.*, et al.* (2009). Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell *139*, 1069-1083.

Linja, M.J., and Visakorpi, T. (2004). Alterations of androgen receptor in prostate cancer. J Steroid Biochem Mol Biol *92*, 255-264.

Liu, J., Ghanim, M., Xue, L., Brown, C.D., Iossifov, I., Angeletti, C., Hua, S., Negre, N., Ludwig, M., Stricker, T.*, et al.* (2009a). Analysis of Drosophila segmentation network identifies a JNK pathway factor overexpressed in kidney cancer. Science *323*, 1218-1222.

Liu, W., Laitinen, S., Khan, S., Vihinen, M., Kowalski, J., Yu, G., Chen, L., Ewing, C.M., Eisenberger, M.A., Carducci, M.A.*, et al.* (2009b). Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. Nat Med *15*, 559-565.

Liu, W., Lindberg, J., Sui, G., Luo, J., Egevad, L., Li, T., Xie, C., Wan, M., Kim, S.T., Wang, Z.*, et al.* (2012). Identification of novel CHD1-associated collaborative alterations of genomic structure and functional assessment of CHD1 in prostate cancer. Oncogene *31*, 3939-3948.

Lukacs, R.U., Memarzadeh, S., Wu, H., and Witte, O.N. (2010). Bmi-1 is a crucial regulator of prostate stem cell self-renewal and malignant transformation. Cell Stem Cell *7*, 682-693.

Lupien, M., Eeckhoute, J.r.m., Meyer, C.a., Wang, Q., Zhang, Y., Li, W., Carroll, J.S., Liu, X.S., and Brown, M. (2008). FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. Cell *132*, 958-970.

Majumder, P.K., Grisanzio, C., O'Connell, F., Barry, M., Brito, J.M., Xu, Q., Guney, I., Berger, R., Herman, P., Bikoff, R.*, et al.* (2008). A prostatic intraepithelial neoplasia-dependent p27 Kip1 checkpoint induces senescence and inhibits cell proliferation and cancer progression. Cancer Cell *14*, 146-155.

Makinen, N., Mehine, M., Tolvanen, J., Kaasinen, E., Li, Y., Lehtonen, H.J., Gentile, M., Yan, J., Enge, M., Taipale, M.*, et al.* (2011). MED12, the mediator complex subunit 12 gene, is mutated at high frequency in uterine leiomyomas. Science *252*.

Mandelbaum, J., Bhagat, G., Tang, H., Mo, T., Brahmachary, M., Shen, Q., Chadburn, A., Rajewsky, K., Tarakhovsky, A., Pasqualucci, L.*, et al.* (2010). BLIMP1 is a tumor suppressor gene frequently disrupted in activated B cell-like diffuse large B cell lymphoma. Cancer Cell *18*, 568-579.

Mani, R.S., Tomlins, S.A., Callahan, K., Ghosh, A., Nyati, M.K., Varambally, S., Palanisamy, N., and Chinnaiyan, A.M. (2009). Induced chromosomal proximity and gene fusions in prostate cancer. Science *326*, 1230.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D.*, et al.* (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med *361*, 1058-1066.

Markert, E.K., Mizuno, H., Vazquez, A., and Levine, A.J. (2011). Molecular classification of prostate cancer using curated expression signatures. Proc Natl Acad Sci *108*, 21276-21281.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

McKie, A.B., Douglas, D.A., Olijslagers, S., Graham, J., Omar, M.M., Heer, R., Gnanapragasam, V.J., Robson, C.N., and Leung, H.Y. (2005). Epigenetic inactivation of the human sprouty2 (hSPRY2) homologue in prostate cancer. Oncogene *24*, 2166-2174.

McMenamin, M.E., Soung, P., Perera, S., Kaplan, I., Loda, M., and Sellers, W.R. (1999). Loss of PTEN expression in paraffin-embedded primary prostate cancer correlates with high Gleason score and advanced stage. Cancer Res *59*, 4291-4296.

Mehra, R., Han, B., Tomlins, S.A., Wang, L., Menon, A., Wasco, M.J., Shen, R., Montie, J.E., Chinnaiyan, A.M., and Shah, R.B. (2007). Heterogeneity of TMPRSS2 gene rearrangements in multifocal prostate adenocarcinoma: molecular evidence for an independent group of diseases. Cancer Res *67*, 7991-7995.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet *11*, 685-696.

Min, J., Zaslavsky, A., Fedele, G., McLaughlin, S.K., Reczek, E.E., De Raedt, T., Guney, I., Strochlic, D.E., Macconaill, L.E., Beroukhim, R.*, et al.* (2010). An oncogene-tumor suppressor cascade drives metastatic prostate cancer by coordinately activating Ras and nuclear factor-kappaB. Nat Med *16*, 286-294.

Mirchandani, D., Zheng, J., Miller, G.J., Ghosh, A.K., Shibata, D.K., Cote, R.J., and Roy-Burman, P. (1995). Heterogeneity in intratumor distribution of p53 mutations in human prostate cancer. Am J Pathol *147*, 92-101.

Miyaki, M., Konishi, M., Kikuchi-Yanoshita, R., Enomoto, M., Igari, T., Tanaka, K., Muraoka, M., Takahashi, H., Amada, Y., Fukayama, M., et al. (1994). Characteristics of somatic mutation of the adenomatous polyposis coli gene in colorectal tumors. Cancer Res 54, 3011-3020.

Mosquera, J.M., Mehra, R., Regan, M.M., Perner, S., Genega, E.M., Bueti, G., Shah, R.B., Gaston, S., Tomlins, S.A., Wei, J.T.*, et al.* (2009). Prevalence of TMPRSS2-ERG fusion prostate cancer among men undergoing prostate biopsy in the United States. Clin Cancer Res *15*, 4706-4711.

Nagai, Y., Kojima, T., Muro, Y., Hachiya, T., Nishizawa, Y., Wakabayashi, T., and Hagiwara, M. (1997). Identification of a novel nuclear speckle-type protein, SPOP. FEBS Lett *418*, 23-26.

Navone, N.M., Labate, M.E., Troncoso, P., Pisters, L.L., Conti, C.J., von Eschenbach, A.C., and Logothetis, C.J. (1999). p53 mutations in prostate cancer bone metastases suggest that selected p53 mutants in the primary site define foci with metastatic potential. J Urol *161*, 304-308.

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M.*, et al.* (2012). The life history of 21 breast cancers. Cell *149*, 994-1007.

Oda, K., Stokoe, D., Taketani, Y., and McCormick, F. (2005). High frequency of coexistent mutations of PIK3CA and PTEN genes in endometrial carcinoma. Cancer Res *65*, 10669-10673.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W.*, et al.* (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet *36*, 1065-1071.

Ouyang, X., DeWeese, T.L., Nelson, W.G., and Abate-Shen, C. (2005). Loss-of-function of Nkx3.1 promotes increased oxidative damage in prostate carcinogenesis. Cancer Res *65*, 6773-6779.

Paez, J.G., Janne, P.A., Lee, J.C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F.J., Lindeman, N., Boggon, T.J.*, et al.* (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science *304*, 1497-1500.

Palanisamy, N., Ateeq, B., Kalyana-Sundaram, S., Pflueger, D., Ramnarayanan, K., Shankar, S., Han, B., Cao, Q., Cao, X., Suleman, K.*, et al.* (2010). Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med *16*, 793-798.

Park, K., Tomlins, S.A., Mudaliar, K.M., Chiu, Y.L., Esgueva, R., Mehra, R., Suleman, K., Varambally, S., Brenner, J.C., MacDonald, T.*, et al.* (2010). Antibody-based detection of ERG rearrangement-positive prostate cancer. Neoplasia *12*, 590-598.

Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L.*, et al.* (2008). An integrated genomic analysis of human glioblastoma multiforme. Science *321*, 1807-1812.

Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V.A., Grunn, A., Messina, M., Elliot, O., et al. (2011). Analysis of the coding genome of diffuse large B-cell lymphoma. Nat Genet 43, 830-837.

Perner, S., Demichelis, F., Beroukhim, R., Schmidt, F.H., Mosquera, J.-M., Setlur, S., Tchinda, J., Tomlins, S.a., Hofer, M.D., Pienta, K.G.*, et al.* (2006). TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Res *66*, 8337-8341.

Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P.C., Svensson, M.A., Kitabayashi, N., Moss, B.J., MacDonald, T.Y.*, et al.* (2011). Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res *21*, 56-67.

Phillips, S.M., Morton, D.G., Lee, S.J., Wallace, D.M., and Neoptolemos, J.P. (1994). Loss of heterozygosity of the retinoblastoma and adenomatous polyposis susceptibility gene loci and in chromosomes 10p, 10q and 16q in human prostate cancer. Br J Urol *73*, 390-395.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordonez, G.R., Bignell, G.R.*, et al.* (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191-196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C.*, et al.* (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184-190.

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol *577*, 67-79.

Prins, G.S., and Putz, O. (2008). Molecular signaling pathways that regulate prostate gland development. Differentiation *76*, 641-659.

Qian, J., Bostwick, D.G., Takahashi, S., Borell, T.J., Herath, J.F., Lieber, M.M., and Jenkins, R.B. (1995). Chromosomal anomalies in prostatic intraepithelial neoplasia and carcinoma detected by fluorescence in situ hybridization. Cancer Res *55*, 5408-5414.

Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H., and Turner, D.J. (2008). A large genome center's improvements to the Illumina sequencing system. Nat Methods *5*, 1005-1010.

Rausch, T., Jones, D.T., Zapatka, M., Stutz, A.M., Zichner, T., Weischenfeldt, J., Jager, N., Remke, M., Shih, D., Northcott, P.A.*, et al.* (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59-71.

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J.P. (2006). GenePattern 2.0. Nat Genet *38*, 500-501.

Rickman, D.S., Soong, T.D., Moss, B., Mosquera, J.M., Dlabal, J., Terry, S., MacDonald, T.Y., Tripodi, J., Bunting, K., Najfeld, V.*, et al.* (2012). Oncogene-mediated alterations in chromatin conformation. Proc Natl Acad Sci *109*, 9083-9088.

Robbiani, D.F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D.J., Chen, H.T., Corcoran, A.E., Nussenzweig, A.*, et al.* (2008). AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. Cell *135*, 1028-1038.

Robbins, C.M., Tembe, W.A., Baker, A., Sinari, S., Moses, T.Y., Beckstrom-Sternberg, S., Beckstrom-Sternberg, J., Barrett, M., Long, J., Chinnaiyan, A.*, et al.* (2011). Copy number and targeted mutational analysis reveals novel somatic events in metastatic prostate tumors. Genome Res *21*, 47-55.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Roychowdhury, S., Iyer, M.K., Robinson, D.R., Lonigro, R.J., Wu, Y.M., Cao, X., Kalyana-Sundaram, S., Sam, L., Balbin, O.A., Quist, M.J.*, et al.* (2011). Personalized oncology through integrative high-throughput sequencing: a pilot study. Sci Transl Med *3*, 111ra121.

Ru, P., Steele, R., Newhall, P., Phillips, N.J., Toth, K., and Ray, R.B. (2012). miRNA-29b suppresses prostate cancer metastasis by regulating epithelial-mesenchymal transition signaling. Mol Cancer Ther *11*, 1166-1173.

Rubin, M.A., Maher, C.A., and Chinnaiyan, A.M. (2011). Common gene rearrangements in prostate cancer. J Clin Oncol *29*, 3659-3668.

Rubin, M.A., Putzi, M., Mucci, N., Smith, D.C., Wojno, K., Korenchuk, S., and Pienta, K.J. (2000). Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. Clin Cancer Res *6*, 1038-1045.

Ruiz, C., Lenkiewicz, E., Evers, L., Holley, T., Robeson, A., Kiefer, J., Demeure, M.J., Hollingsworth, M.A., Shen, M., Prunkard, D.*, et al.* (2011). Advancing a clinically relevant perspective of the clonal nature of cancer. Proc Natl Acad Sci *108*, 12054-12059.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res *20*, 761-770.

Saramaki, O.R., Harjula, A.E., Martikainen, P.M., Vessella, R.L., Tammela, T.L., and Visakorpi, T. (2008). TMPRSS2:ERG fusion identifies a subgroup of prostate cancers with a favorable prognosis. Clin Cancer Res *14*, 3395-3400.

Sasaki, T., Tian, H., Kukita, Y., Inazuka, M., Tahira, T., Imai, T., Yamauchi, M., Saito, T., Hori, T., Hashimoto-Tamaoki, T., et al. (1998). ATM mutations in patients with ataxia telangiectasia screened by a hierarchical strategy. Hum Mutat 12, 186-195.

Sato, K., Qian, J., Slezak, J.M., Lieber, M.M., Bostwick, D.G., Bergstralh, E.J., and Jenkins, R.B. (1999). Clinical significance of alterations of chromosome 8 in high-grade, advanced, nonmetastatic prostate carcinoma. J Natl Cancer Inst *91*, 1574-1580.

Sboner, A., Demichelis, F., Calza, S., Pawitan, Y., Setlur, S.R., Hoshida, Y., Perner, S., Adami, H.O., Fall, K., Mucci, L.A.*, et al.* (2010). Molecular sampling of prostate cancer: a dilemma for predicting disease progression. BMC Med Genomics *3*, 8.

Sharma, A., Yeow, W.S., Ertel, A., Coleman, I., Clegg, N., Thangavel, C., Morrissey, C., Zhang, X., Comstock, C.E., Witkiewicz, A.K.*, et al.* (2010). The retinoblastoma tumor suppressor controls androgen signaling and human prostate cancer progression. J Clin Invest *120*, 4478-4492.

Shen, M.M., and Abate-Shen, C. (2010). Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev *24*, 1967-2000.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.

Shukla, S., Shukla, M., Maclennan, G.T., Fu, P., and Gupta, S. (2009). Deregulation of FOXO3A during prostate cancer progression. Int J Oncol *34*, 1613-1620.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P.*, et al.* (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell *1*, 203-209.

Sircar, K., Yoshimoto, M., Monzon, F.A., Koumakpayi, I.H., Katz, R.L., Khanna, A., Alvarez, K., Chen, G., Darnel, A.D., Aprikian, A.G.*, et al.* (2009). PTEN genomic deletion is associated with p-Akt and AR signalling in poorer outcome, hormone refractory prostate cancer. J Pathol *218*, 505-513.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A.*, et al.* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27-40.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A.*, et al.* (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157-1160.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci *102*, 15545-15550.

Sun, C., Dobi, A., Mohamed, A., Li, H., Thangapazham, R.L., Furusato, B., Shaheduzzaman, S., Tan, S.H., Vaidyanathan, G., Whitman, E.*, et al.* (2008). TMPRSS2-ERG fusion, a common genomic alteration in prostate cancer activates C-MYC and abrogates prostate epithelial differentiation. Oncogene *27*, 5348-5353.

Sun, X., Huang, J., Homma, T., Kita, D., Klocker, H., Schafer, G., Boyle, P., and Ohgaki, H. (2009). Genetic alterations in the PI3K pathway in prostate cancer. Anticancer Res *29*, 1739-1743.

Svensson, M.A., LaFargue, C.J., MacDonald, T.Y., Pflueger, D., Kitabayashi, N., Santa-Cruz, A.M., Garsha, K.E., Sathyanarayana, U.G., Riley, J.P., Yun, C.S.*, et al.* (2011). Testing mutual exclusivity of ETS rearranged prostate cancer. Lab Invest *91*, 404-412.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P.*, et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res *39*, D561-568.

Taylor, B.S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B.S., Arora, V.K., Kaushik, P., Cerami, E., Reva, B.*, et al.* (2010). Integrative genomic profiling of human prostate cancer. Cancer Cell *18*, 11-22.

Tomlins, S.A., Laxman, B., Dhanasekaran, S.M., Helgeson, B.E., Cao, X., Morris, D.S., Menon, A., Jing, X., Cao, Q., Han, B.*, et al.* (2007). Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature *448*, 595-599.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R.*, et al.* (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644-648.

Tomlins, S.A., Rhodes, D.R., Yu, J., Varambally, S., Mehra, R., Perner, S., Demichelis, F., Helgeson, B.E., Laxman, B., Morris, D.S.*, et al.* (2008). The role of SPINK1 in ETS rearrangement-negative prostate cancers. Cancer Cell *13*, 519-528.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*, 562-578.

Tricoli, J.V., Gumerlock, P.H., Yao, J.L., Chi, S.G., D'Souza, S.A., Nestok, B.R., and deVere White, R.W. (1996). Alterations of the retinoblastoma gene in human prostate adenocarcinoma. Genes Chromosomes Cancer *15*, 108-114.

Trotman, L.C., Niki, M., Dotan, Z.A., Koutcher, J.A., Di Cristofano, A., Xiao, A., Khoo, A.S., Roy-Burman, P., Greenberg, N.M., Van Dyke, T.*, et al.* (2003). Pten dose dictates cancer progression in the prostate. PLoS Biol *1*, E59.

True, L., Coleman, I., Hawley, S., Huang, C.Y., Gifford, D., Coleman, R., Beer, T.M., Gelmann, E., Datta, M., Mostaghel, E.*, et al.* (2006). A molecular correlate to the Gleason grading system for prostate adenocarcinoma. Proc Natl Acad Sci U S A *103*, 10991-10996.

Truica, C.I., Byers, S., and Gelmann, E.P. (2000). Beta-catenin affects androgen receptor transcriptional activity and ligand specificity. Cancer Res *60*, 4709-4713.

Ueda, J., Tachibana, M., Ikura, T., and Shinkai, Y. (2006). Zinc finger protein Wiz links G9a/GLP histone methyltransferases to the co-repressor molecule CtBP. J Biol Chem *281*, 20120-20128.

The UniProt Consortium (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res *39*, D214-219.

van Haaften, G., Dalgliesh, G.L., Davies, H., Chen, L., Bignell, G., Greenman, C., Edkins, S., Hardy, C., O'Meara, S., Teague, J., et al. (2009). Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet 41, 521-523.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T.*, et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.

Vasudevan, K.M., Barbie, D.A., Davies, M.A., Rabinovsky, R., McNear, C.J., Kim, J.J., Hennessy, B.T., Tseng, H., Pochanard, P., Kim, S.Y.*, et al.* (2009). AKT-independent signaling downstream of oncogenic PIK3CA mutations in human cancer. Cancer Cell *16*, 21-32.

Vis, A.N., Noordzij, M.A., Fitoz, K., Wildhagen, M.F., Schroder, F.H., and van der Kwast, T.H. (2000). Prognostic value of cell cycle proteins p27(kip1) and MIB-1, and the cell adhesion protein CD44s in surgically treated patients with prostate cancer. J Urol *164*, 2156-2161.

Visakorpi, T., Hyytinen, E., Koivisto, P., Tanner, M., Keinanen, R., Palmberg, C., Palotie, A., Tammela, T., Isola, J., and Kallioniemi, O.P. (1995). In vivo amplification of the androgen receptor gene and progression of human prostate cancer. Nat Genet *9*, 401-406.

Vitari, A.C., Leong, K.G., Newton, K., Yee, C., O'Rourke, K., Liu, J., Phu, L., Vij, R., Ferrando, R., Couto, S.S.*, et al.* (2011). COP1 is a tumour suppressor that causes degradation of ETS transcription factors. Nature *474*, 403-406.

Vocke, C.D., Pozzatti, R.O., Bostwick, D.G., Florence, C.D., Jennings, S.B., Strup, S.E., Duray, P.H., Liotta, L.A., Emmert-Buck, M.R., and Linehan, W.M. (1996). Analysis of 99 microdissected prostate carcinomas reveals a high frequency of allelic loss on chromosome 8p12-21. Cancer Res *56*, 2411-2416.

Voeller, H.J., Truica, C.I., and Gelmann, E.P. (1998). Beta-catenin mutations in human prostate cancer. Cancer Res *58*, 2520-2523.

Wang, Q., Sharma, D., Ren, Y., and Fondell, J.D. (2002). A coregulatory role for the TRAP-mediator complex in androgen receptor-mediated gene expression. J Biol Chem *277*, 42852-42858.

Wang, R., Asangani, I.A., Chakravarthi, B.V., Ateeq, B., Lonigro, R.J., Cao, Q., Mani, R.S., Camacho, D.F., McGregor, N., Schumann, T.E.*, et al.* (2012). Role of transcriptional corepressor CtBP1 in prostate cancer progression. Neoplasia *14*, 905-914.

Wang, X.S., Shankar, S., Dhanasekaran, S.M., Ateeq, B., Sasaki, A.T., Jing, X., Robinson, D., Cao, Q., Prensner, J.R., Yocum, A.K.*, et al.* (2011). Characterization of KRAS Rearrangements in Metastatic Prostate Cancer. Cancer Discov *1*, 35-43.

Watanabe, M., Shiraishi, T., Yatani, R., Nomura, A.M., and Stemmermann, G.N. (1994). International comparison on ras gene mutations in latent prostate carcinoma. Int J Cancer *58*, 174-178.
Williamson, E.a., Wolf, I., O'Kelly, J., Bose, S., Tanosaki, S., and Koeffler, H.P. (2006). BRCA1 and FOXA1 proteins coregulate the expression of the cell cycle-dependent kinase inhibitor p27(Kip1). Oncogene *25*, 1391-1399.

Wiegand, K.C., Shah, S.P., Al-Agha, O.M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M.K., Anglesio, M.S., Kalloger, S.E., et al. (2010). ARID1A mutations in endometriosis-associated ovarian carcinomas. New Engl J Med 363, 1532-1543.

Wu, X., Hepner, K., Castelino-Prabhu, S., Do, D., Kaye, M.B., Yuan, X.J., Wood, J., Ross, C., Sawyers, C.L., and Whang, Y.E. (2000). Evidence for regulation of the PTEN tumor suppressor by a membrane-localized multi-PDZ domain containing scaffold protein MAGI-2. Proc Natl Acad Sci *97*, 4233-4238.

Yang, G., Ayala, G., De Marzo, A., Tian, W., Frolov, A., Wheeler, T.M., Thompson, T.C., and Harper, J.W. (2002). Elevated Skp2 protein expression in human prostate cancer: association with loss of the cyclin-dependent kinase inhibitor p27 and PTEN and with reduced recurrence-free survival. Clin Cancer Res *8*, 3419-3426.

Yang, J., and Weinberg, R.A. (2008). Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Dev Cell *14*, 818-829.

Yardy, G.W., Bicknell, D.C., Wilding, J.L., Bartlett, S., Liu, Y., Winney, B., Turner, G.D., Brewster, S.F., and Bodmer, W.F. (2009). Mutations in the AXIN1 gene in advanced prostate cancer. Eur Urol *56*, 486-494.

Yardy, G.W., and Brewster, S.F. (2005). Wnt signalling and prostate cancer. Prostate Cancer Prostatic Dis *8*, 119-126.

Yegnasubramanian, S., Kowalski, J., Gonzalgo, M.L., Zahurak, M., Piantadosi, S., Walsh, P.C., Bova, G.S., De Marzo, A.M., Isaacs, W.B., and Nelson, W.G. (2004). Hypermethylation of CpG islands in primary and metastatic human prostate cancer. Cancer Res *64*, 1975-1986.

Yu, J., Mani, R.S., Cao, Q., Brenner, C.J., Cao, X., Wang, X., Wu, L., Li, J., Hu, M., Gong, Y.*, et al.* (2010). An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. Cancer Cell *17*, 443-454.

Zhang, C., Wang, L., Wu, D., Chen, H., Chen, Z., Thomas-Ahner, J.M., Zynger, D.L., Eeckhoute, J., Yu, J., Luo, J.*, et al.* (2011). Definition of a FoxA1 Cistrome That Is Crucial for G1 to S-Phase Cell-Cycle Transit in Castration-Resistant Prostate Cancer. Cancer Res *71*, 6738-6748.

Zhang, Q., Shi, Q., Chen, Y., Yue, T., Li, S., Wang, B., and Jiang, J. (2009). Multiple Ser/Thr-rich degrons mediate the degradation of Ci/Gli by the Cul3-HIB/SPOP E3 ubiquitin ligase. Proc Natl Acad Sci *106*, 21191-21196.

Zhou, R., Bonneaud, N., Yuan, C.-X., de Santa Barbara, P., Boizet, B., Schomber, T., Scherer, G., Roeder, R.G., Poulat, F., Berta, P.*, et al.* (2002). SOX9 interacts with a component of the human thyroid hormone receptor-associated protein complex. Nucleic Acids Res *30*, 3245-3252.

Zhuang, M., Calabrese, M.F., Liu, J., Waddell, M.B., Nourse, A., Hammel, M., Miller, D.J., Walden, H., Duda, D.M., Seyedin, S.N.*, et al.* (2009). Structures of SPOP-substrate complexes: insights into molecular architectures of BTB-Cul3 ubiquitin ligases. Mol Cell *36*, 39-50.

Zitzelsberger, H., Engert, D., Walch, A., Kulka, U., Aubele, M., Hofler, H., Bauchinger, M., and Werner, M. (2001). Chromosomal changes during development and progression of prostate adenocarcinomas. Br J Cancer *84*, 202-208.